Time series forecasting by principal covariate regression

Christiaan Heij*, Patrick J.F. Groenen, Dick J. van Dijk
Econometric Institute, Erasmus University Rotterdam

Econometric Institute Report EI2006-37

31-08-2006

Abstract

This paper is concerned with time series forecasting in the presence of a large number of predictors. The results are of interest, for instance, in macroeconomic and financial forecasting where often many potential predictor variables are available. Most of the current forecast methods with many predictors consist of two steps, where the large set of predictors is first summarized by means of a limited number of factors —for instance, principal components— and, in a second step, these factors and their lags are used for forecasting. A possible disadvantage of these methods is that the construction of the components in the first step is not directly related to their use in forecasting in the second step. This motivates an alternative method, principal covariate regression (PCovR), where the two steps are combined in a single criterion. This method has been analyzed before within the framework of multivariate regression models. Motivated by the needs of macroeconomic time series forecasting, this paper discusses two adjustments of standard PCovR that are necessary to allow for lagged factors and for preferential predictors. The resulting nonlinear estimation problem is solved by means of a method based on iterative majorization. The paper discusses some numerical aspects and analyzes the method by means of simulations. Further, the empirical performance of PCovR is compared with that of the two-step principal component method by applying both methods to forecast four US macroeconomic time series from a set of 132 predictors, using the data set of Stock and Watson (2005).

Keywords
principal covariate regression, economic forecasting, dynamic factor models, principal components, distributed lags, iterative majorization

*corresponding author, email address: heij@few.eur.nl

1 Introduction

Econometric modelers face many decisions, as was recently discussed in the special 'Colloquium for ET's 20th Anniversary' issue of this journal by, among others, Hansen (2005), Pesaran and Timmermann (2005), and Phillips (2005). In this paper, we pay attention to one of the basic questions in forecasting, that is, which information should be included in the model. In many cases, observational data are available for a large number of predictor variables that may all help to forecast the variable of interest. To exploit such rich information, one should somehow limit the model complexity, as otherwise the forecasts will suffer from overfitting due to the well-known curse of dimensionality. For instance, if T observations are available for a set of k predictors, then for $k > T$ it is simply impossible to estimate a multiple regression model involving all predictors as separate regressors. If k is large with $k \leq T$, then it is still not advisable to estimate a regression model with all predictors as regressors because the resulting forecasts will have large variance due to overfitting. Better forecasts may be achieved by compressing the information in the predictors somehow and by using a forecast equation containing fewer predictors.

Several methods for forecasting with many predictors have been proposed in the literature. We refer to Stock and Watson (in press) for a survey. For instance, in 'principal component regression' (PCR) the information in the k predictors is summarized by means of a relatively small number of factors (the principal components) and these factors are used as predictors in a low-dimensional multiple regression model. This approach is based on dynamic factor models and is followed, for instance, by Stock and Watson (1999, 2002a,b, 2005) to forecast key macroeconomic variables from large sets of predictor variables. An essential aspect of PCR and similar methods is that they consist of two stages, that is, first the factors are constructed and then the forecast equation is estimated.

The goal of this paper is to analyze a method that combines the two stages of predictor compression and forecasting in a single criterion. This method, called principal covariate regression (PCovR), was proposed by De Jong and Kiers (1992) for multiple regression models. PCovR is a data-based method that does not employ an explicit underlying statistical model. Therefore, we will follow a data analysis approach in our paper and we will not assume a statistical model for the data.

In Heij, Groenen, and Van Dijk (2005), the forecast performance of PCovR and PCR was compared for simple forecast models that employ only the current factors and not their lags. In the current paper, we extend the PCovR method in two respects that are essential for practical applications in economics. The first extension is to allow for preferential predictors, that is, predictors that are always included in the forecast equation, for instance, because of their exceptional forecast power or their economic interpretation. This extension

is relatively straightforward, as the effects of the preferential predictors and of the factors can be estimated in an iterative way. The second extension is to allow for lagged factors in the prediction equation, which is relevant if the effect of the economic predictors is distributed over several periods of time. This extension is much more fundamental, as it requires non-linear estimation methods to respect the condition that the constructed lagged factors all originate from the same underlying factors. We propose an iterative majorization algorithm to estimate the model parameters. We also discuss some numerical aspects of PCovR, that is, the non-convexity of the PCovR criterion function, the issue of initial estimates, and the choice of weight factors in the PCovR criterion.

The forecast performance of PCovR is studied by means of simulation experiments. We investigate various factors that may affect the forecast performance, including the use of preferential predictors, the number of time lags, the number of predictors, and the correlation of the predictors with the variable to be predicted. We make also an empirical comparison of PCovR and PCR by forecasting four key variables of the real US economy (production, income, employment and manufacturing sales) from a set of 132 predictors, using the data set of Stock and Watson (2005). The forecast quality is evaluated by means of the mean squared (several periods ahead, out of sample) forecast error. We consider both single factor and multiple factor models. Model selection is based on the Bayes information criterion, as is common in PCR because of the work of Stock and Watson (1999, 2002a, 2002b, 2005), and also on cross validation methods.

The paper is organized as follows. In Section 2, we formulate the forecasting problem with compressed predictors in more detail and we describe the principal component method. In Section 3, we describe the PCovR method and we present estimation algorithms and discuss some numerical issues, in particular, the choice of PCovR weights. The performance of PCovR under various conditions is analyzed in Section 4 by means of simulation experiments, and Section 5 provides an empirical comparison of PCovR and PCR in macroeconomic forecasting. Section 6 concludes with a brief overview and with some suggestions for further research. Finally, technical results are treated in appendices.

2 Forecasting with compressed predictors

2.1 The forecast model

First we introduce some notation. The observations consist of time series of length T on a variable to be predicted (y) and on a set of predictor variables (X) and preferential predictors (Z; these variables are excluded from X). We will always assume that the constant term is excluded both from X and from Z. Let k be the number of predictors and k_z the number of preferential predictors, then y is a $T \times 1$ vector, X is a $T \times k$ matrix and Z is a $T \times k_z$

matrix.

The idea is to summarize the information in the k variables X by means of p factors F, with p (much) smaller than k. Here F is a $T \times p$ matrix consisting of linear combinations of the X variables, so that

$$F = XA$$

for some $k \times p$ matrix A. These factors are used, together with the preferential predictors, to forecast y by means of a distributed lag model. If q lags of F and r lags of Z are incorporated in the model then the (one-step-ahead) forecast equation for y_{T+1} at period T is written as

$$\hat{y}_{T+1} = \alpha + \sum_{j=0}^{q} f_{T-j}\beta_j + \sum_{j=0}^{r} z_{T-j}\gamma_j. \tag{1}$$

Here α is the constant term, β_j are $p \times 1$ vectors ($j = 0, \ldots, q$), γ_j are $k_z \times 1$ vectors ($j = 0, \ldots, r$), and $f_{T-j} = x_{T-j}A$ where x_t denotes the $1 \times k$ vector of observations on the predictors at time t. The (multi) h-step-ahead forecast equation has the same structure, replacing $\hat{y}_{T+1}$ in (1) by $\hat{y}_{T+h}$. In the sequel, we mostly consider the case $h = 1$, but the methods are easily extended for $h > 1$. In the empirical application in Section 5, the forecast horizon is $h = 12$ months.

To apply the forecast model (1) in practice, we should choose the structure parameters (p, q, r) and estimate the parameters $(A, \alpha, \beta_0, \ldots, \beta_q, \gamma_0, \ldots, \gamma_r)$ of the forecast equation. In this paper, we pay most attention to the estimation of the parameters for a given set of structure parameters. However, in the simulation experiments in Section 4, we consider the effects of misspecification of (p, q, r), and in Sections 4 and 5 we consider the forecast performance if the structure parameters (p, q, r) are selected by the Bayes information criterion or by cross validation.

2.2 Two-step principal component regression (PCR)

In this section, we briefly describe the method of principal component regression (PCR) to estimate the forecast equation (1). We refer to Stock and Watson (1999, 2002a,b, 2005) for more details and for applications in macroeconomic forecasting.

The PCR method consists of two estimation steps. In the first step, A is estimated by means of principal components. That is, the p factors are obtained by minimizing the squared Frobenius norm $||X - \hat{X}||^2$ under the restriction that $\hat{X}$ has rank p. The squared Frobenius norm of a matrix is simply the sum of squares of all elements of the matrix. The X-variables should be standardized to prevent scale effects. For instance, each column (variable) of X is scaled to have zero mean and unit norm.

The estimates A can be obtained from the singular value decomposition (SVD) of X. More precisely, let $X = USV'$ be an SVD of X where the singular values in the diagonal

matrix S are listed in decreasing order. Then $\hat{X} = U_p S_p V_p'$ where U_p and V_p consist respectively of the first p columns of U and V and where S_p is the $p \times p$ diagonal matrix with the p largest (non-zero) singular values of X on the diagonal. Define the $k \times p$ matrix $A = V_p S_p^{-1}$ and $p \times k$ matrix $B = S_p V_p'$, then $\hat{X} = XAB$ and (A, B) provides the minimizing solution of

$$||X - XAB||^2, \qquad \text{with } A \;\; k \times p \quad \text{and} \quad B \;\; p \times k. \tag{2}$$

It is easily checked that the factors $F = XA$ satisfy $F'F = A'X'XA = I_p$, so that the p factors in F are scaled and mutually orthogonal. The factors $F = XA$ are called the principal components of X.

In the second step, the parameters $(\alpha, \beta_0, \ldots, \beta_q, \gamma_0, \ldots, \gamma_r)$ in (1) are estimated by least squares (OLS), for given values of A. Let $F = XA$ with corresponding lagged matrices $F(-1), \ldots, F(-q)$, and let $Z(-1), \ldots, Z(-r)$ be the lagged matrices of Z. Then the second step corresponds to minimizing

$$||y - \alpha - \sum_{j=0}^{q} F(-j)\beta_j - \sum_{j=0}^{r} Z(-j)\gamma_j||^2, \tag{3}$$

where some initial observations —that is, the first $\max(q, r)$ ones— should be dropped because of missing observations for the lagged terms of F and Z.

Summarizing, PCR consists of the (SVD) minimization (2) followed by the (OLS) minimization (3). In the next section, we consider a method that integrates these two steps by minimizing a single criterion function.

3 Principal covariate regression (PCovR)

3.1 Introduction

In this section, we consider a method for forecasting with many predictors that combines the two stages of predictor compression and estimating the parameters of the forecast equation into a single criterion. This method is called 'Principal Covariate Regression' (PCovR) and was proposed by De Jong and Kiers (1992) within the framework of multiple regression models. We first describe 'standard' PCovR and we discuss extensions with preferential predictors in Section 3.2 and with lagged factors in Section 3.3.

In PCovR, the parameters are estimated by minimizing a weighted average of the forecast errors (3) (without preferential predictors Z and without lags of F, so that $q = 0$) and of the predictor compression errors (2). For given weights $w_1 > 0$ and $w_2 > 0$ and for given number of factors p, the criterion to be minimized is

$$f(A, B, \alpha, \beta) = w_1||y - \alpha - XA\beta||^2 + w_2||X - XAB||^2, \tag{4}$$

where the $T \times p$ matrix $F = XA$ consists of p factors that compress the predictor information in the $T \times k$ matrix X. As before, A is a $k \times p$ matrix of rank p, B is a $p \times k$ matrix, α is a scalar and β is a $p \times 1$ vector. Clearly, if (A, B, α, β) is an optimal set of coefficients then $(AR, R^{-1}B, \alpha, R^{-1}\beta)$ is also optimal for every invertible $p \times p$ matrix R. Therefore, A may be chosen such that $F'F = A'X'XA = I_p$, as $p \leq \text{rank}(X)$. With this restriction, the parameters are identified up to an orthogonal transformation R, that is, with $R'R = I_p$.

The vector norm in (4) is the Euclidean norm and the matrix norm is the Frobenius norm. To prevent scaling effects of the X-variables, we will assume that all variables —that is, all columns of X— are scaled to have mean zero and norm one. Further, because only the relative weight w_1/w_2 is of importance, we consider weights of the form

$$w_1 = \frac{w}{||y||^2}, \qquad w_2 = \frac{1-w}{||X||^2}, \tag{5}$$

with $0 \leq w \leq 1$. For scaled predictor data there holds $||X||^2 = k$, where k is the number of predictor variables. The user has to choose the PCovR weight w, balancing the objectives of good predictor compression for X (for w small) and good (in-sample) fit for y (for w large). The parameter w should be chosen between 0 and 1, because otherwise the criterion (4) becomes unbounded and has no optimal solution. If the weight w tends to 0 then PCovR converges to Principal Components (step 1 in PCR), and if w tends to 1 then PCovR converges to OLS.

The minimization of (4) is a nonlinear —in fact, bilinear— optimization problem, because of the product terms $A\beta$ and AB. The estimates can be obtained by means of two SVD's, see Heij, Groenen, and Van Dijk (2005). We refer to De Jong and Kiers (1992) for further background on PCovR in multiple regression models. The next two subsections discuss two extensions that are needed in time series forecasting, namely, the inclusion of preferential predictors and of lags in the forecast equation (1).

3.2 PCovR with preferential predictors

In many cases, one wishes to include some of the variables explicitly in the forecast equation, and not indirectly via the constructed factors. For instance, to predict y one may wish to use lagged values of y or variables closely related to y, and possibly also some variables suggested by (economic) theory. These preferential predictors are denoted by Z in the forecast equation (1). If we assume that there are no lags, so that $q = r = 0$ in (1), then the PCovR criterion with preferential predictors is given by

$$f(A, B, \alpha, \beta, \gamma) = w_1||y - \alpha - XA\beta - Z\gamma||^2 + w_2||X - XAB||^2, \tag{6}$$

where Z is a $T \times k_z$ matrix and γ a $k_z \times 1$ vector.

A simple, alternating least squares method to estimate $(A, B, \alpha, \beta, \gamma)$ is the following. In Step 1, regress y on a constant and Z to get initial estimates a of α and c of γ. In Step 2, use the residuals $(y - a - Zc)$ and X to estimate (A, B, β) by standard PCovR as described in the foregoing section. These steps can be iterated, where at each new iteration updates of the estimates of α and γ are obtained by regressing $y - X\hat{A}b$ on a constant and Z, where $\hat{A}$ and b are the estimates of A and β obtained in the previous iteration. At each iteration and at both steps, the criterion value $f(A, B, \alpha, \beta, \gamma)$ in (6) decreases and therefore converges, since it is bounded from below by zero.

3.3 PCovR with lagged factors

The methods discussed so far are too limited to deal with time series data, as the effect of the predictors X and Z on y will often be distributed over several time periods. Such distributed effects may occur, for instance, in macroeconomic applications where adjustments may be relatively slow. In such situations it is useful to incorporate lags in the forecast equation, as in (1) with q lags of X and r lags of Z.

Lags of preferential predictors can be incorporated simply by extending the matrix Z with additional columns for the lagged terms, so that Z in (6) is replaced by $Z_r = [Z \;\; Z(-1) \;\; \ldots \;\; Z(-r)]$. In principle, one possible solution to add lagged factor terms is to extend also the original predictor matrix X with lagged terms, so that X in (6) is replaced by $X_q = [X \;\; X(-1) \;\; \ldots \;\; X(-q)]$. The advantage is that the parameters can be estimated in a relatively simple way. However, this method has three important disadvantages. The first objection is of a practical nature. The motivation for the predictor compression is that the number of predictors k in X is relatively large. However, this dimensionality problem is magnified if we use X_q, as this matrix has $k(q + 1)$ columns instead of k. The second objection is related to the interpretation of the constructed factors. If $F = XA$, then at time t the p factor values $F_t = X_t A$ consist of linear combinations of the values of the original predictor variables that are all observed at the same time t. However, factors $F_q = X_q A$ consist of mixtures of variables measured at different points in time, which makes it more difficult to interpret the factors. The third objection is that a large number of predictors requires the use of relatively small values for the weight factor w, as will be explained in Section 3.5. As X_q has much more columns than X, the weight should be decreased accordingly. In this case, the relative importance of the (in-sample) fit of y in (4), (6) decreases, which may be a disadvantage in forecasting y.

Because of these considerations, we consider the PCovR criterion that is based directly on the forecast equation (3). The criterion is given by $f = f(A, B, \alpha, \beta_0, \ldots, \beta_q, \gamma_0, \ldots, \gamma_r)$

with

$$\begin{aligned} f &= w_1||y-\alpha-\sum_{j=0}^{q}F(-j)\beta_j-\sum_{j=0}^{r}Z(-j)\gamma_j||^2+w_2||X-FB||^2 \\ &= w_1||y-\alpha-\sum_{j=0}^{q}(XA)(-j)\beta_j-\sum_{j=0}^{r}Z(-j)\gamma_j||^2+w_2||X-XAB||^2, \end{aligned} \tag{7}$$

where $F = XA$ are the factors to be constructed and $F(-j) = (XA)(-j)$ and $Z(-j)$ are lags of respectively $F = F(0) = XA$ and $Z = Z(0)$. We include a constant term α in (7) to allow for possible non-zero sample means of the variables and their lags.

An advantage of this method is that the constructed factors have the usual interpretation of linear combinations of economic variables measured at the same point in time. A computational disadvantage is that the minimization of (7) can no longer be solved by SVD's because of the implicit parameter restrictions for the parameters of X, $X(-1), \ldots, X(-q)$ in (7). The resulting minimization problem can be solved by an iterative approximation method known as 'iterative majorization'. This approach is discussed in general terms in the next subsection, and further details are given in Appendix A.1.

3.4 Iterative majorization for PCovR with lagged factors

The idea to solve the nonlinear minimization problem (7) is to approximate the criterion function by a simpler (quadratic) one that can be solved by SVD, and to iterate the approximation to get closer and closer to the minimum of the original criterion function (7). In this section, we give a general outline of the method, and details of the algorithm are given Appendix A.1. For more background on iterative majorization we refer to Kiers (1990, 2002), De Leeuw (1994), Heiser (1995), Lange, Hunter and Yang (2000), and Borg and Groenen (2005).

Note that the non-linearity is only due to the parameter matrix A. Indeed, if A is fixed then $F = XA$ and its lags are known, so that the parameters $(\alpha, \beta_0, \ldots, \beta_q, \gamma_0, \ldots, \gamma_r)$ in (7) can be estimated by regressing y on a constant and F and Z and their lags, and (each column of) B can be estimated by regressing (each column of) F on X. These regression estimates depend on A, and if we substitute the estimates in (7) then PCovR boils down to minimizing a non-linear function $f(A)$ of the $k \times p$ matrix A, which is normalized so that $A'X'XA = I_p$.

The idea of majorization is to find a (local) minimum of f in an iterative way, where at each iteration the objective function f is replaced by a computationally simpler function g, the so-called majorizing function, with the following two properties: $f(A) \leq g(A)$ for all A, so that g majorizes f, and $f(\overline{A}) = g(\overline{A})$ at the current estimate $\overline{A}$. If g is minimal for $A = A^*$, then it follows that $f(A^*) \leq g(A^*) \leq g(\overline{A}) = f(\overline{A})$ and that $f(A^*) < f(\overline{A})$ if

$g(A^*) < g(\overline{A})$. By means of this iterative majorization we obtain a sequence of estimates of A with monotonically decreasing function values $f(A)$. The estimates converge to a local minimum of f under suitable regularity conditions.

As is shown in Appendix A.1, in each iteration the PCovR criterion function can be majorized by a function $g(A) = v - \text{trace}(A'VA)$ where the scalar v and the $k \times k$ (semi-definite positive) matrix V are computed from the observed data (y, X, Z) and from the estimate of A in the previous iteration. The minimization of g, that is, the maximization of $\text{trace}(A'VA)$ under the condition $A'X'XA = I_p$, can be solved by SVD on a $k \times p$ matrix, which provides a new estimate of A. The computations can be simplified even further to the SVD of a $p \times p$ matrix in each iteration step. This simplification is computationally attractive, as in practice the number of factors p will be much smaller than the number of predictors k. Details of the algorithm are given in Appendix A.1.

We mention some numerical aspects of minimizing the PCovR criterion. The criterion function (4) is not convex in the parameters (A, B, α, β). Therefore, the criteria (6) and (7) are also not convex. This is shown in Appendix A.2. Further, the criterion functions may have several local minima. As the majorization method in practice converges to a local minimum it is not guaranteed to arrive at a global minimum. Therefore, the choice of initial estimates may of importance, as this choice may affect the attained (local or global) minimum. Experiments with numerous random starts and with starts based on PCR did not reveal different local minima, as always the same forecasts of y and approximations of X were obtained, independent of the chosen initial estimates. Therefore, it seems that PCovR criteria like (7) do not involve serious problems with local minima, so that we will be satisfied with the estimates obtained by the iterative majorization algorithm.

3.5 Choice of weight to prevent overfitting

The PCovR criterion functions in the foregoing sections seek to find a balance between the two objectives to forecast y and to compress the predictive information in X by means of a small number of factors. The weights w_1 and w_2 in (4), (6) and (7) are defined in (5) in terms of the PCovR weight $0 < w < 1$. In this section, we show that this weight should be chosen sufficiently small in order to prevent overfitting of y and that the upper bound on w decreases with the number of predictors k and with the number of factors p. For example, if the number of predictors exceeds the number of observations so that $k \geq T$, then —in the generic case that the $T \times k$ matrix X has full row rank— we can reconstruct any $T \times 1$ vector y by means of $Fb = XAb$. By letting w approach to 1, the PCovR criterion value in (4), (6) and (7) decreases towards 0, so that PCovR leads to overfitting of y in such situations.

To analyze this in more detail, we will assume in this section that the data are generated by a factor model. We refer to Bai and Ng (2002), Boivin and Ng (2006), Forni et al. (2000,

2003), and Stock and Watson (2002a) for further background on factor models. We can derive an upper bound for the weight w for these models, by requiring that a specifically constructed overfitting estimate does not provide a lower PCovR criterion value than the data generating process itself. That is, the bound is derived by preventing a specific overfitting estimate. Of course, other overfitting estimates should also be excluded, so that the derived upper bound on w is not necessarily the tightest one.

In a factor model, the observed variables (y, X) are generated by underlying factors F by means of

$$y = F\beta + \varepsilon \qquad \text{and} \qquad X = F\Lambda + V. \tag{8}$$

Here F is a $T \times p_0$ matrix of (unobserved) factors, β is a $p_0 \times 1$ vector, Λ is a $p_0 \times k$ matrix of factor loadings, V and ε are respectively a $T \times k$ matrix and a $T \times 1$ vector with independent error terms, and (F, ε, V) are mutually independent. The p_0 factors are scaled to have zero mean and unit covariance matrix $E(F_t' F_t) = I_{p_0}$, and all observed variables —y and the columns of X— are scaled to have zero mean and unit variance. Although the model equations in (8) are written in non-dynamic form, lagged effects can be modelled by including lagged factor values in the factor matrix F and by postulating a dynamic model for the factors.

Let x_i, λ_i and v_i be the $T \times 1$ vectors consisting respectively of the i-th column of X, Λ and V, so that $x_i = F\lambda_i + v_i$. Define $\rho^2_{x_i F}$ and ρ^2_{yF} as respectively the squared correlation between x_i and F and between y and F. As y, x_i and F all have unit variance and (F, ε, v_i) are independent, it follows that

$$\rho^2_{yF} = 1 - \sigma^2_\varepsilon, \qquad \rho^2_{x_i F} = 1 - \sigma^2_{v_i}.$$

Now suppose that we model the data by means of PCovR with p factors, where p may be equal to p_0 or not. In a sense, the 'true' model is the one by which the data are generated (if $p \geq p_0$) or the approximating model (if $p < p_0$) obtained by taking only the first p principal factors (that is, the ones explaining most of the variance). If we substitute the factors and parameters of the data generating process into the PCovR criterion and approximate sample averages by population means, we get the (asymptotic) approximations

$$\begin{aligned} \frac{||y - F\beta||^2}{||y||^2} &\approx \sigma^2_\varepsilon = 1 - \rho^2_{yF}(p), \\ \frac{||X - F\Lambda||^2}{||X||^2} &\approx \frac{1}{k}\sum_{i=1}^{k} \sigma^2_{v_i} = \frac{1}{k}\sum_{i=1}^{k}\Big(1 - \rho^2_{x_i F}(p)\Big), \end{aligned}$$

where $\rho^2_{yF}(p)$ and $\rho^2_{x_i F}(p)$ are the squared correlations of the first p factors of F with respectively y and x_i, with $\rho^2_{yF}(p) = \rho^2_{yF}$ and $\rho^2_{x_i F}(p) = \rho^2_{x_i F}$ if $p \geq p_0$. To simplify the notation, we define the average correlation $\rho^2_{XF}(p)$ of all k predictors with the first p factors

by

$$\rho^2_{XF}(p) = \frac{1}{k}\sum_{i=1}^{k}\rho^2_{x_iF}(p),$$

so that $(1/k)\sum_{i=1}^{k}(1-\rho^2_{x_iF}(p)) = 1-\rho^2_{XF}(p)$. The PCovR criterion value (4), (5) obtained for the parameters of the data generating process is

$$w\frac{||y-F\beta||^2}{||y||^2} + (1-w)\frac{||X-F\Lambda||^2}{||X||^2} \approx w(1-\rho^2_{yF}) + (1-w)\Big(1-\rho^2_{XF}(p)\Big).$$

We wish to prevent possible overfitting of y in situations where the number of predictors k is relatively large as compared to the number of observations T. For this purpose, we consider one particular option for overfitting, that is, to use a single factor to fit y with a maximal number of zero errors and to use the other $(p-1)$ factors to approximate the predictors X. If $k < T$ then we can create (at least) k errors in $y-\hat{y} = y - XAb$ with value zero, with corresponding (sample) R-squared $R^2_{y\hat{y}} \approx 1-(T-k)/T = k/T$ and with $||y-\hat{y}||^2/||y||^2 \approx 1-R^2_{y\hat{y}}$. If $k \geq T$ then we can recreate $y = XAb$ without errors and with $R^2_{y\hat{y}} = 1$. The remaining $(p-1)$ factors to approximate X can be chosen, for instance, as the $(p-1)$ principal factors of the data generating process with resulting fit $||X-\hat{X}||^2/||X||^2 \approx 1-\rho^2_{XF}(p-1)$. The resulting overfitted model has an (asymptotic) PCovR criterion value

$$w(1-R^2_{y\hat{y}}) + (1-w)\Big(1-\rho^2_{XF}(p-1)\Big).$$

To prevent this kind of overfitting, this (asymptotic) PCovR criterion value should be larger than the one obtained for the parameters of the data generating process. This condition is equivalent to

$$w(\rho^2_{yF}-R^2_{y\hat{y}}) + (1-w)\Big(\rho^2_{XF}(p)-\rho^2_{XF}(p-1)\Big) > 0,$$

which provides the following upper bound for the weight w:

$$w < \frac{\rho^2_{XF}(p)-\rho^2_{XF}(p-1)}{(R^2_{y\hat{y}}-\rho^2_{yF}) + (\rho^2_{XF}(p)-\rho^2_{XF}(p-1))}.$$

As discussed before, $R^2_{y\hat{y}} \approx \min(1, k/T)$, but the population correlations in (9) are of course unknown. In Appendix A.3, we prove that the correlation between X and F with p factors can be approximated by

$$\rho^2_{XF}(p) \approx \frac{\sum_{i=1}^{p} s_i^2}{\sum_{i=1}^{k} s_i^2},$$

where s_i^2 are the squared singular values of X, in decreasing order. This means that $\rho^2_{XF}(p) - \rho^2_{XF}(p-1) \approx s_p^2/\sum_{i=1}^{k} s_i^2$. Further, to stay on the safe side, we take $\rho^2_{yF} = 0$, which gives the following upper bound for w.

$$w < \frac{s_p^2/\sum_{i=1}^{k} s_i^2}{\min(1, k/T) + s_p^2/\sum_{i=1}^{k} s_i^2}. \tag{9}$$

If $p > 1$ then it is also possible to spend more factors to fit y, which will increase the fit if $k < T$. For instance, if $p_m \leq p$ is the largest value with $p_m k \leq T$ then we can use p_m factors to fit y and $(p - p_m)$ factors to fit X. In this way, we can derive another upper bound for w. However, as is proven in Appendix A.3, this bound is not smaller than that in (9), so that we will use (9) as upper bound.

It should be noted that the bound (9) is a rough approximation, as we only considered one single option for overfitting. Even lower bounds may be necessary in practice to exclude other types of overfitting. Our main message is that overfitting can only be prevented if the weight is chosen small enough. In particular, the bound may become very small for rich predictor sets with k relatively large as compared to T and with slowly decaying singular values. In the extreme case without decay, that is, if all singular values of X are equal, the bound becomes

$$w < \frac{1}{1 + \min(k, k^2/T)} \to 0 \quad \text{if} \quad \frac{k^2}{T} \to \infty. \tag{10}$$

As $w = 0$ corresponds to PCR, this result shows that PCovR becomes asymptotically equivalent to PCR in such situations.

4 Simulation experiment for PCovR

4.1 Data generating process

We illustrate the PCovR method of Sections 3.3 and 3.4 by simulating data from simple data generating processes (DGP's). The specification of the DGP and of the employed PCovR model are varied to investigate the forecast performance under different conditions. For instance, we vary the number of predictor variables and the lag structure of the DGP, the correlations between the observed variables, the lag structure of the PCovR model, and the choice of the PCovR weight. In all cases, the purpose is to forecast the dependent variable y one-step-ahead on the basis of observed past data on y and on a set of predictors X and, possibly, on a set of preferential predictors Z. The forecast quality is measured by the mean squared forecast error (MSE) of y over a number of simulation runs.

More specifically, we consider the following data generating process:

$$y_t = x^*_{t-L_1} + \gamma z_{t-L_2} + \varepsilon_t. \tag{11}$$

Here y, x^* and z are observed scalar variables and ε is (unobserved) white noise. The predictors x^* and z are all independent white noise processes with zero mean and unit variance. The information set at the forecast moment consists of observations over the time interval $t = 1, \ldots, 100$ for y_t and over the interval $t = 1, \ldots, 101$ for a set of k variables X (including x^*) and for the single variable z, and the purpose is to forecast y_t at $t = 101$.

We consider the following specifications of the DGP. The number of predictors k is either 10, 40 or 100; note that in the last case the number of predictors is equal to the number of observations. The lag L_1 of x^* is either 1 or 5. The preferential predictor z may be absent (for $\gamma = 0$, which we will sometimes indicate by writing $L_2 = -1$), and if z is present (with $\gamma \neq 0$) then the lag L_2 is either 1 or 5. Further, we consider values of 0.5 and 0.9 for the squared partial correlations ρ^2_{yx} and ρ^2_{yz}, defined as the squared correlation respectively between $(y_t - \gamma z_{t-L_2})$ and $x^*_{t-L_1}$ and between $(y_t - x^*_{t-L_1})$ and z_{t-L_2}. The desired value of ρ^2_{yx} is achieved by choosing the variance σ^2_ε of ε appropriately, and ρ^2_{yz} is achieved by choosing γ appropriately. More precisely, $\rho^2_{yx} = \text{var}(x^*)/\text{var}(x^* + \varepsilon) = 1/(1 + \sigma^2_\varepsilon)$ so that

$$\sigma^2_\varepsilon = \frac{1 - \rho^2_{yx}}{\rho^2_{yx}}, \tag{12}$$

and $\rho^2_{yz} = \text{var}(\gamma z)/\text{var}(\gamma z + \varepsilon) = \gamma^2/(\gamma^2 + \sigma^2_\varepsilon)$ so that

$$\gamma^2 = \sigma^2_\varepsilon \frac{\rho^2_{yz}}{1 - \rho^2_{yz}} = \frac{\rho^2_{yz}(1 - \rho^2_{yx})}{\rho^2_{yx}(1 - \rho^2_{yz})}. \tag{13}$$

For instance, for $\rho^2_{yx} = 0.5$ and $\rho^2_{yz} = 0.5$ we take $\sigma^2_\varepsilon = 1$ and $\gamma = 1$, and for $\rho^2_{yx} = 0.9$ and $\rho^2_{yz} = 0.9$ we take $\sigma^2_\varepsilon = 1/9$ and $\gamma = 1$.

With three options for k, six for the lag structure (L_1, L_2), and four for the correlations $(\rho^2_{yx}, \rho^2_{yz})$, this gives in total seventy-two configurations. However, in the twenty-four combinations with $\gamma = 0$, the correlation between y and z is by definition zero so that the distinction by means of ρ^2_{yz} drops out, reducing the number of configurations by twelve. That is, we consider in total sixty specifications of the DGP.

4.2 Forecast models and evaluation

The considered PCovR models have either one or two factors, so that $p = 1$ or $p = 2$. The maximum lags of the factor and the preferential predictor are chosen to be equal, with $L = q = r = 1$ or 5. This set-up implies that some of the models are over-specified, with too many factors, with a superfluous preferential predictor, or with too large lags. Other models are under-specified, with too small lags. Five values are considered for the PCovR weight w, namely, (0.0001, 0.01, 0.1, 0.5, 0.9).

With two options for p, two for the lags (q, r) and five for w, this gives in total twenty PCovR models. However, for some DGP's with many predictors we do not consider the largest weights, because of the bound (9) derived in Section 3.5. For this simulation, we can easily compute the (theoretical) singular values needed in (9), as all variables in the $T \times k$ matrix X (with $k \leq T$) are independent with equal variance, so that the covariance matrix of X has k identical eigenvalues and we can take $s^2_i = 1$ in (9). As $k \leq T$, the bound then

becomes

$$w < \frac{1/k}{(k/T) + (1/k)} = \frac{1}{1 + k^2/T}.$$

As $T = 100$, the bound gives $w < 0.5$ for $k = 10$, $w < 0.06$ for $k = 40$, and $w < 0.01$ for $k = 100$. For the sake of illustration, we will use a maximum weight of $w = 0.9$ for $k = 10$, of $w = 0.5$ for $k = 40$, and of $w = 0.1$ for $k = 100$, to check for possible overfitting in these situations.

For each DGP we perform one thousand simulation runs. The data of each run are used to compute one-step-ahead forecasts of y for each of the twenty PCovR models. The models are compared by the MSE, that is,

$$\text{MSE}_j = \frac{1}{1000} \sum_{i=1}^{1000} \frac{(y_i - \hat{y}_{ij})^2}{\sigma_\varepsilon^2} \tag{14}$$

where j denotes the employed PCovR model, y_i is the actual value of y at the forecast time $t = 101$ in the i-th simulation run, and $\hat{y}_{ij}$ is the value forecasted by method j in the i-th simulation run. The squared forecast error $(y_i - \hat{y}_{ij})^2$ is divided by the error variance σ_ε^2, as this provides a natural benchmark for the forecast errors that would be obtained if the DGP (11) were estimated perfectly. So we should expect that MSE > 1 in all cases, and values close to 1 indicate near optimal forecasts.

The variance of the dependent variable y depends on the DGP. Therefore, to facilitate the interpretation of the reported MSE values, we also report the MSE for the model-free 'zero-prediction' $\hat{y} = 0$, which is equal to $\text{var}(y)/\sigma_\varepsilon^2$. We call this the relative variance of y, denoted by $\text{rvar}(y)$. The variance of y in (11) is $(1 + \gamma^2 + \sigma_\varepsilon^2)$, and with the expressions for σ_ε^2 and γ^2 obtained in (12) and (13) we get

$$\text{rvar}(y) = \frac{\text{var}(y)}{\sigma_\varepsilon^2} = 1 + \frac{1}{\sigma_\varepsilon^2} + \frac{\gamma^2}{\sigma_\varepsilon^2} = 1 + \frac{\rho_{yx}^2}{1 - \rho_{yx}^2} + \frac{\rho_{yz}^2}{1 - \rho_{yz}^2}.$$

For each PCovR model, the parameters are estimated by minimizing the criterion function (7) in Section 3.4 by means of iterative majorization. The employed tolerance to stop the PCovR iterations is that the relative decrease in the criterion function (7) is smaller than 10^{-6}. However, as the number of considered DGP and model combinations is large (around a thousand, resulting in around a million simulations), we limit the number of PCovR iterations to a maximum of one hundred in each instance.

4.3 Forecast results

The MSE results of the simulations are shown in Table 1. The DGP's are shown in rows and the PCovR models in columns. To limit the size of the table, we report only the outcomes for thirty-six of the sixty considered DGP's (i.e., with identical correlations of y with x and z, equal to 0.5 or 0.9) and for four of the five considered PCovR weights w (i.e., 10^{-4}, 0.1,

0.5 and 0.9) (more detailed results are available on request). The results for $w = 0.01$ are very close to those for $w = 10^{-4}$ for $k = 10$ and $k = 40$, with MSE differences of at most 3%. The MSE's for $w = 10^{-4}$ are also very close to those of PCR which are, therefore, not reported separately. Further, in most cases the results for $(\rho^2_{yx}, \rho^2_{yz}) = (0.5, 0.9)$ are similar to those for (0.5,0.5), and the results for (0.9,0.5) are similar to those for (0.9,0.9).

As the MSE in (14) is measured relative to the best (DGP) predictor, all MSE values are larger than 1. The column 'rvar(y)' shows the MSE of the 'zero-prediction', so that MSE values between 1 and 'rvar(y)' show the forecast gain of PCovR. The model is correctly specified if $p = 1$ and $L = L_1 = L_2$, whereas the model is under-specified if $L < \max(L_1, L_2)$ and over-specified if $p = 2$ or $L > \min(L_1, L_2)$. As was discussed in the previous section, for some DGP's the PCovR model is not estimated for $w = 0.9$ and $w = 0.5$ so that the columns for these weights are not complete.

<< Table 1 to be included around here. >>

First we discuss the results for the relatively simple case of $k = 10$ predictors. In this case, the forecasts are in general best for large PCovR weights, that is, for $w = 0.5$ and $w = 0.9$. The MSE values for $p = 1$ and $w = 0.9$ vary roughly between 1.2 and 1.4 for all DGP's, provided that the lags in the PCovR model are not too small. Of course, if the model lag L is smaller than the DGP factor lag L_1, then the predictors x do not help in forecasting y because the relevant predictor $x^*_{t-L_1}$ is uncorrelated with all considered predictors $x_{i,t-L}$ with $L < L_1$. As expected, the MSE increases for over-specified models, i.e., if the lag of the model is larger than that of the DGP or if the model has $p = 2$ factors instead of $p = 1$ (note that the DGP can be forecasted by a single factor, namely x^*). For instance, for the DGP with lags $L_1 = L_2 = 1$, the MSE for $w = 0.5$ and 0.9 is around 1.2 for the PCovR model with correct specification $(p, L) = (1, 1)$, whereas it is around 1.4 for $(p, L) = (1, 5)$, 1.3 for $(p, L) = (2, 1)$, and 1.7 for $(p, L) = (2, 5)$.

For the case of $k = 40$ predictors, most MSE values are somewhat larger than for $k = 10$. For DGP's with large predictor lag $L_1 = 5$, PCovR produces relatively better forecasts for $\rho^2_y = 0.9$ than for $\rho^2_y = 0.5$. Note that, for $L = 5$, the task is to find the DGP predictor $x^*_{t-L_1}$ in the set of 240 possible predictors consisting of the forty predictors and their five lagged values. The optimal weight w depends on the DGP. For instance, for the simple DGP with $(L_1, L_2, \rho^2_y) = (1, -1, 0.9)$, the best choice is $w = 0.5$, whereas for the more complex DGP with $(L_1, L_2, \rho^2_y) = (5, 1, 0.5)$ it is $w = 10^{-4}$. Further note that the MSE for $w = 0.5$ is relatively large in many cases, which is consistent with the result in Section 3.5 that suggests an upper bound (for $k = 40$ and $T = 100$) of $w < 0.06$. In general, the optimal weight tends to be smaller for DGP's with more lags and smaller correlations. In the next section, we will consider data-based methods to select the weight.

Finally, for the case of $k = 100$ predictors, acceptable results are only obtained for the smallest considered weight $w = 10^{-4}$. Even for $w = 0.01$ most of the forecasts are useless, as the MSE is much larger than that of the naive zero-prediction, and for $w > 0.01$ the forecasts are even worse. This finding is in line with the bound $w < 0.01$ for this case, but the results for $w = 0.01$ indicate that this bound may still be too large to prevent overfitting. PCovR with $w = 0.0001$ does not suffer from overfitting, but it does also not succeed in exploiting the information in the hundred predictors x. Indeed, the zero-prediction is only beaten for DGP's with a preferential predictor ($L_2 = 1$ or 5), whereas PCovR performs worse than this simple benchmark for DGP's without preferential predictor ($L_2 = -1$).

As was stated in Section 4.2, the number of majorization iterations to estimate each PCovR model is limited to at most one hundred. For $k = 10$, this iteration bound is nearly never reached, except for $(p, L, w) = (2, 5, 0.9)$, that is, for the case of an over-specified model with large PCovR weight. In general, the iteration bound is restrictive only in overfitting situations where the PCovR weight w is larger than the upper bound of Section 3.5. We investigated for several of these cases whether the early stop of the iterations caused the large MSE values, but this does not seem to be the case. If no upper bound on the number of iterations is imposed in these overfitting cases, then the PCovR majorization algorithm may take a very large number of iterations (up to several thousands), but the forecasts do not improve.

4.4 Model selection

The results in the foregoing section show that the optimal PCovR weight w depends on the DGP and on the applied forecast model. Therefore, we now consider the performance of PCovR if the structure parameters of the forecast model are not fixed a priori but are selected by using the Bayes information criterion (BIC) or cross validation (CV). That is, for each simulated data set, a set of models with different weights w, number of factors p and lag lengths q and r are estimated, and the model parameters (p, q, r, w) are selected by minimizing BIC or the cross validation MSE. In addition, we consider also the selection of (p, q, r) for a given PCovR weight w. The set of considered models is larger than that of Section 4.3, as the models have $p = 1$, 2, or 3 factors with possibly different lags q and r, which are both chosen from the set $\{-1, 0, 1, 2, 5, 6, 10\}$. Here $q = -1$ ($r = -1$) means that the forecast equation (1) does not contain any factor (preferential predictor). The considered DGP's have lags -1, 1 or 5, so that the model set contains models with the correct DGP lags as well as models with too small or too large lags. The considered PCovR weights are $\{10^{-4}, 0.01, 0.1, 0.5, 0.9\}$. This gives in total 735 models.

For each simulated data set, the task is to select one of the 735 models. For BIC this is done as follows. For each model (p, q, r, w), the PCovR criterion (7) is minimized with

$k = p$ factors. This delivers fitted values of y_t on the time interval $t \leq 100$, by substituting the constructed PCovR factors f_t and the PCovR coefficients of (7) into the right-hand-side of the forecast equation (1). Let the corresponding residual variance be s^2, then the BIC value of the model is $\log(s^2) + d\log(T_e)/T_e$, where $d = p(q+1) + r + 2$ is the number of parameters in (1) and $T_e = 100 - \max(q, r)$ is the effective number of observations in (1). The model with the lowest BIC value is selected to forecast y on $t = 101$.

As an alternative, we consider five-fold cross validation to select the model. In this case, the observation sample with $t \leq 100$ is split into five equally large parts. Each of the five parts is used as a validation sample, in which case the data of the other four parts are used to estimate PCovR models and to compute the corresponding MSE in forecasting the validation sample. The selected model is the one with the smallest average MSE over the five validation samples. This selected model is estimated once more, now using all observations $t \leq 100$ in the minimization of (7), and the resulting model is used to forecast y on $t = 101$.

The MSE results are in Table 2, and Table 3 contains information on the selected models and weights. We report the results only for a subset of all considered DGP's, the same as in Table 1. The columns show the MSE if BIC or CV are used for a given weight w, and also if in addition this weight is also selected by BIC or CV. For BIC, it turns out that the selected weight is always minimal ($w = 10^{-4}$), although this weight is often not the one giving the best forecasts. The reason is that the fit of y contributes the term $\log(s^2)$ to BIC, which in our simulations is relatively small as compared to the term $pq\log(T_e)/T_e$ related to the number of factor lags. As the results in Table 3 show, BIC prefers to choose q (too) small, which goes at the expense of larger values of s^2. This lack of attention to the fit of y corresponds to a small PCovR weight w in (4) and (5). In short, BIC is not suited well to choose w, and Table 2 shows that cross validation works much better in this respect.

<< Tables 2 and 3 to be included around here. >>

Table 2 shows that CV has lower MSE than BIC for nearly all DGP's with $k = 10$ and $k = 40$. We mention some results of interest, consecutively for $k = 10$, 40 and 100. For $k = 10$, the differences between BIC and CV are not so large for DGP's with factor lag $L_1 = 1$, but if $L_1 = 5$ then CV is far better than BIC for all weights. Full CV to choose all model parameters (p, q, r, w) works reasonably well in all cases. Table 3 shows that the average selected weight is relatively large, with averages of around 0.65 for $\rho_y^2 = 0.5$ and 0.85 for $\rho_y^2 = 0.9$. Table 3 shows also that CV tends to choose much fewer factors p than BIC, that CV performs much better in selecting the factor lag q (BIC mostly chooses q much smaller than the DGP lag L_1), and that BIC tends to be better in selecting the lag r of the preferential predictor.

The results for $k = 40$ are roughly similar. Again, CV works better than BIC in most cases, and full CV gives good MSE results. The selected weight is smaller than for $k = 10$, with averages smaller than 0.1 for $\rho_y^2 = 0.5$ and between 0.3 and 0.5 for $\rho_y^2 = 0.9$. As compared to BIC, CV again chooses fewer factors and it is better in selecting q, whereas BIC is better in selecting r. For $k = 100$, we only consider $w = 10^{-4}$. CV and BIC give comparable MSE's in this case. CV performs again better in the selection of p and q, although it has problems in choosing q if the DGP lag is $L_1 = 5$, and BIC is better in selecting r.

5 Empirical comparison of PCovR and PCR

5.1 Data and forecast design

We use the data set of Stock and Watson (2005) for an empirical comparison of PCovR with PCR. In a series of papers, Stock and Watson (1999, 2002a, 2002b) applied PCR to forecast key macroeconomic variables in the US for different forecast horizons, using monthly observations of a large set of predictors. Here, we consider forecasts of four variables of the real economy, that is, industrial production, employment, income, and manufacturing sales, with forecast horizons of six, twelve and twenty-four months. As predictors we take 128 variables of the 132 in Stock and Watson (2005) (we exclude their four regional housing starts variables because of some missing observations, but we include the series of total housing starts). The monthly data are available over the period 1959.01 to 2003.12 and are transformed to get stationary series without serious outliers. We refer to Stock and Watson (2005) for a more detailed description of the used data set.

Our purpose is to illustrate the results of PCovR in forecasting the four mentioned series and to compare the outcomes with those of PCR. We follow the same forecast set-up as in Stock and Watson (2002b), where the four variables are forecasted from a different set of predictors over a shorter observation interval. We use the forecast model (1), which is called a diffusion index model with autoregressive terms and lags (DI-AR-Lag) in Stock and Watson (2002b). They also consider restricted models without autoregressive terms and factor lags, but here we will restrict the attention to the DI-AR-Lag model. The forecasted variable is the (annualized) growth of the economic variable of interest over the considered forecast horizon, and the preferential predictor z_t is the one-month growth of the economic variable over the last month. Simulated out-of-sample forecasts are computed over the period 1970.01 till 2003.12$-h$, where h is the forecast horizon, and the forecast quality is measured by the MSE of the corresponding $408 - h$ forecasts over this period. This MSE is expressed relative to an AR benchmark, that is, the forecast model (1) with $q = -1$ so that it contains only a constant and z_T and its lags as predictors.

We consider the same model set as in Stock and Watson (2002b), with $1 \leq p \leq 4$ factors, $0 \leq q \leq 2$ factor lags, and $-1 \leq r \leq 5$ autoregressive lags, where $r = -1$ means that also the zero-order lag term z_T is missing in (1). This setting defines a set of 84 models. For PCovR, we consider the weights w in the set $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9\}$, giving in total 504 models. The results for weights $w < 0.01$ are close to those of PCR and are therefore not reported. Apart from multiple factor models (with $p \leq 4$), we consider also single factor models (with $p = 1$).

For a fair comparison of PCovR with PCR, we optimize the set-up for PCR in three respects. Firstly, we use the PCR algorithm proposed in Heij, Van Dijk and Groenen (2006), because this improves the PCR forecasts somewhat as compared to the method of Stock and Watson (2002b). Secondly, we use a moving window of fifteen years (180 observations) to estimate the forecast model, instead of the expanding window used in Stock and Watson (2002b). This choice reduces the MSE of PCR by around $5 - 10\%$ for most variables and forecast horizons (here we do not provide further details, which are available on request). Thirdly, we applied PCR both with BIC and with CV, and it turned out that BIC gives the lowest MSE in the far majority of cases. Therefore, we use BIC to select the model parameters (p, q, r), both for PCR and for PCovR. The PCovR weight w is chosen afterwards by CV, as follows. Let (p_w, q_w, r_w) be the model selected by BIC for each of the six considered weights w, then w is selected by CV on the resulting set of six models. We use a CV algorithm with buffers, as proposed by Burman, Chow and Nolan (1994) and Racine (2000), to reduce the correlation between the estimation and validation samples. More precisely, the window of 180 observations is split into three parts, a validation sample of 36 observations, two buffers around this sample of 12 observations, and an estimation sample consisting of the remaining 120 observations.

In the above set-up, with $T = 180$ observations and $k = 128$ predictors, the rough upper bound (10) on the weight gives $w < 0.01$. This bound means that a direct application of PCovR, with such small weights, will give results that are close to PCR. As we aim for a comparison of both methods, larger weights are of more interest, which is possible by reducing the number of predictors. For simplicity, we use a method that does not affect PCR, namely, by replacing the predictors by their leading principal components. More precisely, at time t we replace the 180×132 matrix X_t with the observations of the 132 predictors on the estimation interval $[t-179, t]$ by the $180 \times \hat{k}$ matrix $\hat{X}_t$, defined as follows. Let $X_t = U_t S_t V_t'$ be an SVD of X_t, then $\hat{X}_t = US$ where U consists of the first $\hat{k}$ columns of U_t and S consists of the first $\hat{k}$ rows and columns of S_t. We choose $\hat{k} = 10$, so that the upper bound (10) becomes $w < 0.65$. PCR with $p \leq 4$ is not affected by this change of predictors, as the four leading principal components of X_t and $\hat{X}_t$ are identical, that is, the first four columns of U_t. The choice of $\hat{k} = 10$ is somewhat arbitrary, as smaller values give

a larger upper bound (10) and larger values retain more of the original predictor variance. PCR uses at most four principal components, which (measured over the full sample period) account for roughly one-third of the total predictor variance. Ten principal components account for around one half of this predictor variance, whereas accounting for two-thirds of the variance requires twenty components, in which case the bound (10) becomes $w < 0.3$. We take $\hat{k} = 10$ to allow for somewhat larger weights. In practice, the joint choice of T and $\hat{k}$ is of interest, but we will not analyze this any further here.

5.2 Forecast results

The forecast results of PCovR and PCR are summarized in Table 4 (for single factor models with $p = 1$) and Table 5 (for multiple factor models with $p \leq 4$). The table rows show the four forecasted variables and the three considered forecast horizons, and the columns show the applied forecast method. The MSE's of PCR and PCovR are measured relative to the MSE of the AR benchmark model. For PCovR, results are shown both for the (data-based, time dependent) CV weights w and for the (a posteriori, time independent) optimal choice of a fixed weight w. This (a posteriori) optimal choice is not feasible in practice, but it serves as a benchmark to evaluate the quality of the CV weights. The last four columns show the forecast gains of PCovR as compared to PCR over the full sample and over three sub-samples, two (70-80 and 81-91) with 132 months and one (92-03) with $144 - h$ months where h is the forecast horizon.

<< Tables 4 and 5 and Figure 1 to be included around here.>>

Table 4 shows that the PCR and PCovR single factor models both improve much on the AR benchmark and that PCovR provides better forecasts than PCR in nearly all cases. The results of PCovR with CV weight (column 'PCovR') are, as expected, somewhat worse than with a posteriori chosen fixed weight (column 'PCovR*'), but cross validation works reasonably well. The gains are, on average, largest for production and sales, and smallest for income. Averaged over the four variables and over the full sample period, the MSE gain of PCovR as compared to PCR is 18.6% for $h = 6$, 23.4% for $h = 12$, and 26.3% for $h = 24$. The gain tends to be larger for longer forecast horizon and for early sample periods.

A comparison of Tables 4 and 5 shows that multiple factor models provide better forecasts than single factor models. Further, PCR improves relatively more than PCovR by adding extra factors. Stated otherwise, the first PCovR factor is a better predictor than the first PCR factor, but additional PCR factors add more to the forecast power than additional PCovR factors. For a horizon of $h = 6$ months, the full sample gains are 7.9% for employment, 4.9% for sales, 0.5% for income and -4.7% for production, with an average

gain of 2.1%. The forecasts are worse for $h = 12$, with an average loss of 3.2%. The most consistent gains are for $h = 24$, with an average gain of 6.0%. PCovR does not succeed in forecasting the production series any better than PCR, but forecast gains can be achieved for the other three series. This result means that, with the set-up chosen in Section 5.1, none of the two methods is uniformly better.

Figure 1 shows the MSE gains when evaluated over forecast intervals starting in 1970.01 and ending at varying times, ranging from 1975.01 till the end of the full sample period. The gains are most persistent for $h = 24$, with largest gains in initial periods and with losses sometimes later on. As concerns the prediction of the production series, for $h = 6$ PCovR looses much on PCR in initial periods and it relatively improves in later periods, whereas for $h = 24$ it gains much in initial periods but looses afterwards.

5.3 Further comparison of PCovR and PCR

Table 6 summarizes some results on the structure of the selected multiple factor forecast models of PCovR and PCR and on the series of forecast errors of both methods. The rows in this table are the same as in Tables 4 and 5. The first columns of the table show the mean values of the parameters (p, q, r, w) of the selected forecast models. On average, PCovR uses somewhat fewer factors than PCR, whereas the average lags q and r are comparable for both methods. The cross validation weight w does not vary much across the considered variables and forecast horizons and lies mostly between 0.2 and 0.4.

The last columns in Table 6 show three statistics of the forecast errors, that is, the mean value, the mean absolute value, and the standard deviation. PCR and PCovR have roughly the same bias and variance. The bias tends to be somewhat smaller for PCovR, and the variance is smaller for PCR if $h = 12$ but it is smaller for PCovR if $h = 24$. Both methods have a comparable mean absolute error. This result indicates that outliers do not play an important role, which could be expected because the data in Stock and Watson (2005) have been treated for outliers.

We also performed the test of Diebold and Mariano (1995), with robust standard errors, to examine whether PCovR provides a significantly lower MSE than PCR. For multiple factor models, the differences are not significant when evaluated over the full sample period. However, PCovR is significantly better than PCR for some of the variables over some of the subperiods 1970-1980, 1981-1991 and 1991-2003, six times at a 10% significance level and four times at a 5% level. Further, PCovR is never significantly worse than PCR for any of the series over any of the subperiods or over the full sample period, at a 10% significance level. As could be expected from the results in Table 4, PCovR is often significantly better than PCR for single factor models.

<< **Table 6 to be included around here.** >>

6 Conclusion

In this paper, we considered the Principal Covariate Regression (PCovR) method. This method estimates factors that approximate the predictors X and the dependent variable y by minimizing a criterion that consists of a weighted average of the squared errors for y and those for X. We presented an iterative estimation method for the resulting nonlinear estimation problem and discussed the choice of weights in the criterion function. The forecast quality of PCovR under various circumstances was analyzed by means of a simulation study. Further, an empirical comparison of PCovR and PCR was made by one-year-ahead forecasts of four macroeconomic variables (production, income, employment and manufacturing sales). The results show that PCovR can be a valuable tool in forecasting.

We conclude by mentioning some extensions that are of possible interest. In the empirical application, the model structure (p, q, r) was selected by BIC. Another option is to use forecast oriented selection methods, for instance, cross validation methods. Further, we showed that, to prevent overfitting, the PCovR weight should be small if the number of predictors is large. It is of interest to develop methods for regularization of PCovR to prevent overfitting for larger weights. As an alternative, first a subset of the predictors can be selected and then PCovR can be applied (with larger weights) on this smaller set.

References

[1] Bai, J., and S. Ng (2002), Determining the number of factors in approximate factor models, *Econometrica* 70, pp. 191-221.

[2] Borg, I., and P.J.F. Groenen (2005), *Modern Multidimensional Scaling*, 2-nd ed., New York, Springer.

[3] Boivin, J, and S. Ng (2006), Are more data always better for factor analysis?, *Journal of Econometrics* 132, pp. 169-194.

[4] Burman, P, E. Chow and D. Nolan (1994), A cross-validatory method for dependent data, *Biometrika* 81, pp. 351-358.

[5] De Leeuw, J. (1994), Block relaxation algorithms in statistics, in H.H. Bock, W. Lenski and M.M. Richter (eds.), *Information Systems and Data Analysis*, Berlin, Springer Verlag, pp. 308-324.

[6] De Jong, S., and H.A.L. Kiers (1992), Principal covariate regression, *Chemometrics and Intelligent Laboratory Systems* 14, pp. 155-164.

[7] Diebold, F.X., and Mariano, R.S. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-263.

[8] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000), The generalized dynamic factor model: identification and estimation, *Review of Economics and Statistics* 82, pp. 540-554.

[9] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2003), Do financial variables help forecasting inflation and real activity in the euro area?, *Journal of Monetary Economics* 50, pp. 1243-1255.

[10] Hansen, B.E. (2005), Challenges for econometric model selection, *Econometric Theory* 21, pp. 60-68.

[11] Heij, C., P.J.F. Groenen and D.J. van Dijk (2005), Forecast comparison of principal component and principal covariate regression, *Research Report 2005-28*, Econometric Institute Rotterdam. Submitted.

[12] Heij, C., P.J.F. Groenen and D.J. van Dijk (2006), Improved construction of diffusion indexes for macroeconomic forecasting, *Research Report 2006-03*, Econometric Institute Rotterdam. Submitted.

[13] Heiser, W.J. (1995), Convergent computation by iterative majorization: Theory and applications, in W.J. Krzanowski (ed.), *Recent Advances in Descriptive Multivariate Analysis*, Oxford, Oxford University Press, pp. 157-189

[14] Kiers, H.A.L. (2002), Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems, *Computational Statistics and Data Analysis* 41, pp. 157-170.

[15] Kiers, H.A.L. (1990), Majorization as a tool for optimizing a class of matrix functions, *Psychometrika* 55, pp. 417-428.

[16] Lange, K., D.R. Hunter and I. Yang (2000), Optimization transfer using surrogate objective functions, *Journal of Computational and Graphical Statistics* 9, pp. 1-20.

[17] Pesaran, H, and A. Timmermann (2005), Real-time econometrics, *Econometric Theory* 21, pp. 212-231.

[18] Phillips, P.C.B. (2005), Automated discovery in econometrics, *Econometric Theory* 21, pp. 3-20.

[19] Racine, J. (2000), Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation, *Journal of Econometrics* 99, pp. 39-61.

[20] Stock, J.H., and M.W. Watson (1999), Forecasting inflation, *Journal of Monetary Economics* 44, pp. 293-335.

[21] Stock, J.H., and M.W. Watson (2002a), Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association* 97, pp. 1167-1179.

[22] Stock, J.H., and M.W. Watson (2002b), Macroeconomic forecasting using diffusion indixes, *Journal of Business and Economic Statistics* 20, pp. 147-162.

[23] Stock, J.H., and M.W. Watson (2005), Implications of dynamic factor models for VAR analysis, *Working Paper*.

[24] Stock, J.H., and M.W. Watson (in press), Forecasting with many predictors, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, North-Holland, Amsterdam (to appear).

Appendix

This appendix contains proofs of some results and an algorithm for PCovR with lagged factors. Interested readers can contact the authors for Matlab routines of this algorithm.

Appendix A.1 (Section 3.4)
Majorization algorithm for PCovR with lagged factors

We present an iterative majorization algorithm for the minimization of (7) in Section 3.4. We assume that the estimation interval contains no 'missing' observations. Such missing observations may occur in cross validation if the validation sample is taken somewhere in the middle of the sample. The algorithm should be slightly adjusted in that case to incorporate the selection of data for the estimation sample, but for simplicity we do not discuss the details here. First we derive the various required steps in (a) to (e), and then we summarize this by means of a step-by-step algorithm in (f).

(a) Simplification of the optimization problem

For given values of $(A, B, \beta_0, \ldots, \beta_q)$, the parameters $(\alpha, \gamma_0, \ldots, \gamma_r)$ are simply obtained by regressing the residuals $(y - \sum_{j=0}^{q}(XA)(-j)\beta_j)$ on a constant and Z and its r lags. Next, the parameters $(A, B, \beta_0, \ldots, \beta_q)$ can be updated by applying the majorization algorithm below on the residuals $(y - \alpha - \sum_{j=0}^{r} Z(-j)\gamma_j)$ and X. Therefore, in what follows we can restrict the attention to (7) without constant term and without Z and its lags, so that we wish to minimize

$$f(A, B, \beta) = w_1||y - \sum_{j=0}^{q}(XA)(-j)\beta_j||^2 + w_2||X - XAB||^2. \tag{15}$$

(b) Notation

We introduce some notation, in addition to the notation used in Section 3.4. Because of the lags of F and Z in (7), the relevant vector for y in (7) contains the observations on the time interval $[m+1, T]$ where $m = \max(q, r)$. By X we denote from now on the 'q-past extended' $(T - m + q) \times k$ matrix with the values of the predictors over the interval $[m - q + 1, T]$. Let S_j be the j-th $(T - m) \times (T - m + q)$ shift matrix consisting of $(q - j)$ columns of zeros, followed by the identity matrix, followed by j columns of zeros, that is, $S_j = [O_{(T-m)\times(q-j)} \quad I_{(T-m)\times(T-m)} \quad O_{(T-m)\times j}]$. Then the criterion (15) can be written more explicitly as

$$f(A, B, \beta) \quad = \quad w_1||y - \sum_{j=0}^{q} S_j XA\beta_j||^2 + w_2||X - XAB||^2$$

$$= \quad w_1||y - \sum_{j=0}^{q} S_j c_j||^2 + w_2||X - XAB||^2, \tag{16}$$

where $c_j = XA\beta_j$. Then $c = (c_0', \ldots, c_q')'$ is a $(q+1)(T-m+q)\times 1$ vector, $R = [S_0\ S_1\ \ldots\ S_q]$ is a $(T-m)\times(q+1)(T-m+q)$ matrix, $\sum_{j=0}^{q} S_j c_j = Rc$, and

$$||y - \sum_{j=0}^{q} S_j c_j||^2 = (y-Rc)'(y-Rc) = ||y||^2 + c'R'Rc - 2c'R'y. \tag{17}$$

Here $R'R$ is the $(q+1)(T-m+q)\times(q+1)(T-m+q)$ matrix with $(T-m+q)\times(T-m+q)$ blocks $S_h'S_i$ on the (h,i)-th block position $(h, i = 0, \ldots, q)$.

(c) Construction of majorizing function

Let $\theta = (A, B, \beta)$, with $\beta = (\beta_0', \ldots, \beta_q')'$, be the parameter vector and let $\overline{\theta} = (\overline{A}, \overline{B}, \overline{\beta})$ be a given set of parameter values, that is, initial values or the values obtained at a previous iteration. To apply the idea of majorization we need to construct a function $g(\theta)$ with the properties that $f(\theta) \le g(\theta)$ for all θ and $f(\overline{\theta}) = g(\overline{\theta})$. To find a suitable function g, we will expand the function f in (16) and (17) explicitly.

The (non-zero) eigenvalues of $R'R$ are the same as those of $RR' = \sum_{j=0}^{q} S_j S_j' = (q+1)I_{(T-m)\times(T-m)}$, that is, $R'R$ has $(T-m)$ eigenvalues equal to $(q+1)$ and the other eigenvalues are all zero, so that $(q+1)I_{(q+1)(T-m+q)\times(q+1)(T-m+q)} - R'R$ is a positive semi-definite matrix. Let $\overline{c}_j = X\overline{A}\overline{\beta}_j$ correspond to the current parameter estimates and write $\lambda = (q+1)$, then $\lambda(c-\overline{c})'(c-\overline{c}) - (c-\overline{c})'R'R(c-\overline{c}) \ge 0$ so that

$$c'R'Rc \le \lambda c'c + \lambda\overline{c}'\overline{c} - 2\lambda c'\overline{c} - \overline{c}'R'R\overline{c} + 2c'R'R\overline{c}. \tag{18}$$

We now exploit the fact that the p factors $F = XA$ may always be chosen so that $F'F = A'X'XA = I_{p\times p}$. With this choice, the term $c'c$ in (18) simplifies, as $c'c = \sum \beta_j' A'X'XA\beta_j = \sum \beta_j'\beta_j$. If we substitute (17) and (18) in (16), it follows that

$$\begin{aligned}
f(A,B,\beta) &\le w_1\Big(||y||^2 + \lambda c'c + \lambda\overline{c}'\overline{c} - 2\lambda c'\overline{c} - \overline{c}'R'R\overline{c} + 2c'R'R\overline{c} - 2c'R'y\Big) \\
&\quad + w_2||X - XAB||^2 \\
&= w_1\Big(\lambda\sum\beta_j'\beta_j - 2c'(\lambda\overline{c} - R'R\overline{c} + R'y)\Big) + w_2||X - XAB||^2 + a \\
&= w_1\Big(\lambda\sum\beta_j'\beta_j - 2\sum c_j'(\lambda\overline{c}_j + S_j'(y - R\overline{c}))\Big) + w_2||X - XAB||^2 + a \\
&= w_1\Big(\lambda\sum\beta_j'\beta_j - 2\sum c_j'u_j\Big) + w_2||X - XAB||^2 + a \\
&= w_1\Big(\lambda\sum\beta_j'\beta_j - 2\sum\beta_j'A'X'u_j\Big) + w_2||X - XAB||^2 + a \\
&= g(A,B,\beta),
\end{aligned}$$

where $a = w_1(||y||^2 + \lambda\overline{c}'\overline{c} - \overline{c}'R'R\overline{c})$ and $u_j = \lambda\overline{c}_j + S_j'(y - R\overline{c})$ are fixed, that is, obtained from the data and the previous estimates and not dependent on the values of (A, B, β). The function g majorizes f (as the above derivation shows) and it also satisfies the condition that $f(\overline{\theta}) = g(\overline{\theta})$ at $\overline{\theta} = (\overline{A}, \overline{B}, \overline{\beta})$, as for $c = \overline{c}$ the inequality in (18) becomes an equality.

The essential simplification of $g(A, B, \beta)$ as compared to $f(A, B, \beta)$ is (18), where the term $c'R'Rc$, which is quadratic in the bilinear terms $c_j = XA\beta_j$, is approximated by an expression that is linear in c_j and quadratic in β_j (as $c'c = \sum \beta_j'\beta_j$). That is, the term $c'R'Rc$ in (17), which is of order four in the parameters, is approximated by a second-order expression.

(d) Minimization of majorizing function

The function g can be minimized, as follows. For given values of A with $A'X'XA = I_{p\times p}$, the optimal values of B and β_j are simply obtained by regression, with solutions

$$\hat{B} = \Big((XA)'XA\Big)^{-1}(XA)'X = A'X'X, \qquad \hat{\beta}_j = \frac{1}{\lambda}A'X'u_j.$$

If we substitute these values in g, it remains to minimize the function of A defined by $g(A, \hat{B}, \hat{\beta})$. We use the following facts, where tr denotes the trace of a matrix.

$$\lambda\sum \hat{\beta}_j'\hat{\beta}_j - 2\sum \hat{\beta}_j'A'X'u_j = -\frac{1}{\lambda}\sum u_j'XAA'X'u_j = -\frac{1}{\lambda}\mathrm{tr}(A'X'(\sum u_ju_j')XA),$$

$$\begin{aligned} ||X - XA\hat{B}||^2 &= \mathrm{tr}(X - XA\hat{B})'(X - XA\hat{B}) = \mathrm{tr}\Big(X'X + \hat{B}'A'X'XA\hat{B} - 2X'XA\hat{B}\Big) \\ &= \mathrm{tr}\Big(X'X + \hat{B}'\hat{B} - 2X'XA\hat{B}\Big) = \mathrm{tr}\Big(X'X - X'XAA'X'X\Big) \\ &= \mathrm{tr}(X'X) - \mathrm{tr}(A'X'XX'XA). \end{aligned}$$

It follows that

$$\begin{aligned} g(A, \hat{B}, \hat{\beta}) &= \mathrm{tr}\Big(-\frac{w_1}{\lambda}A'X'(\sum u_ju_j')XA - w_2(A'X'XX'XA)\Big) + w_2\mathrm{tr}(X'X) + a \\ &= -\mathrm{tr}(A'VA) + v, \end{aligned}$$

where $V = (w_1/\lambda)X'(\sum u_ju_j')X + w_2X'XX'X$ and $v = a + w_2\mathrm{tr}(X'X)$ are constant, that is, independent of the value of A. It follows that the minimization of g boils down to the minimization of

$$-\mathrm{tr}(A'VA), \qquad V = \frac{w_1}{\lambda}X'(\sum u_ju_j')X + w_2X'XX'X, \tag{19}$$

under the condition that $A'X'XA = I_{p\times p}$. Let $D = (X'X)^{1/2}A$ so that $D'D = I_{p\times p}$, and let $W = (X'X)^{-1/2}V(X'X)^{-1/2}$ have eigendecomposition $W = Q\Sigma Q'$, then $\mathrm{tr}(A'VA) = \mathrm{tr}(D'Q\Sigma Q'D)$ where the $k \times p$ matrix $Q'D$ satisfies $D'QQ'D = I_{p\times p}$. The optimal choice is to take $Q'D$ equal to $(I_{p\times p} \;\; O)'$, so that D consists of the first p columns of Q, which we denote by Q_p, and the resulting function value is $-\mathrm{tr}(A'VA) = -\sum_{i=1}^p \sigma_i$ where $(\sigma_1, \ldots, \sigma_p)$ are the p largest eigenvalues of W. Therefore, the optimal value of A is $\hat{A} = (X'X)^{-1/2}Q_p$, with resulting minimal function value

$$g(\hat{A}, \hat{B}, \hat{\beta}) = w_1\Big(||y||^2 + \lambda\overline{c}'\overline{c} - \overline{c}'R'R\overline{c}\Big) + w_2\mathrm{tr}(X'X) - \sum_{i=1}^p \sigma_i.$$

Because of the majorizing properties of g, it follows that

$$f(\hat{A}, \hat{B}, \hat{\beta}) \leq g(\hat{A}, \hat{B}, \hat{\beta}) \leq g(\overline{A}, \overline{B}, \overline{\beta}) = f(\overline{A}, \overline{B}, \overline{\beta}).$$

This majorization can now be applied iteratively, returning to (c) to construct a new majorizing function, using the new estimates $\hat{\theta} = (\hat{A}, \hat{B}, \hat{\beta})$ instead of the old estimates $\overline{\theta} = (\overline{A}, \overline{B}, \overline{\beta})$.

(e) Additional majorization for numerical simplification

From a numerical point of view, the main step in the minimization problem in (d) is the eigendecomposition of the $k \times k$ matrix $W = (X'X)^{-1/2}V(X'X)^{-1/2}$. As the number of predictors k may be large this may be a relatively time consuming operation, especially because this eigendecomposition should be performed in each iteration of the majorization algorithm. We now describe an additional iterative majorization method where the high-dimensional eigendecomposition of W is approximated by iterations involving the eigendecomposition of $p \times p$ matrices where, in practice, p is much smaller than k.

Let $G = (X'X)^{1/2}A$, then the minimization of $-\text{tr}(A'VA)$ in (19) subject to the condition $A'X'XA = I_{p\times p}$ is equivalent to the minimization of $h(G) = -\text{tr}(G'WG)$ subject to the condition $G'G = I_{p\times p}$. Define $\overline{G} = (X'X)^{1/2}\overline{A}$ where $\overline{A}$ is the current estimate of A. As V in (19) is positive semi-definite, the same holds true for W, so that $\text{tr}(G-\overline{G})'W(G-\overline{G}) \geq 0$ and hence

$$h(G) = -\text{tr}(G'WG) \leq -2\text{tr}(G'W\overline{G}) + \text{tr}(\overline{G}'W\overline{G}) = h^*(G).$$

As $h(\overline{G}) = h^*(\overline{G})$ this means that h^* can be used as majorizing function for h. The minimization of h^* is equivalent to

$$\text{maximize } \text{tr}(G'W\overline{G}), \text{ subject to } G'G = I_{p\times p}. \tag{20}$$

Let $M = \overline{G}'W^2\overline{G}$, then M is a $p \times p$ positive semi-definite matrix (we assume that M is positive definite, as will be true generically for reasonable values of p, but the steps below are easily adjusted in case M is only positive semi-definite). Let $M = V_m\Sigma_m V_m'$ be an eigendecomposition of M, and define $U_m = W\overline{G}V_m\Sigma_m^{-1/2}$. We will prove that $\hat{G} = U_mV_m'$ solves (20), so that the majorizing function h^* can be minimized by an eigendecomposition of the $p \times p$ matrix M. This result gives the claimed numerical simplification by iterative majorization to minimize (19).

It remains to prove that $\hat{G} = U_mV_m'$ solves (20). This result follows, for instance, from the solution of the so-called orthogonal Procrustean problem in Borg and Groenen (2005, Section 20.2). However, for the sake of completeness we give an explicit proof. As

$$U_m'U_m = \Sigma_m^{-1/2}V_m'\overline{G}'W^2\overline{G}V_m\Sigma_m^{-1/2} = \Sigma_m^{-1/2}V_m'MV_m\Sigma_m^{-1/2} = I_{p\times p},$$

it follows that $\hat{G}'\hat{G} = V_m U_m' U_m V_m' = I_{p\times p}$. Further, as $W\overline{G} = U_m \Sigma_m^{1/2} V_m'$ it follows that

$$\mathrm{tr}(G'W\overline{G}) = \mathrm{tr}(G'U_m\Sigma_m^{1/2}V_m') = \mathrm{tr}(V_m'G'U_m\Sigma_m^{1/2}).$$

If we take $\hat{G} = U_m V_m'$, then this expression delivers the value $\mathrm{tr}(\Sigma_m^{1/2})$ in (20), and we should prove that for any other value of G with $G'G = I_{p\times p}$ there holds $\mathrm{tr}(V_m'G'U_m\Sigma_m^{1/2}) \leq \mathrm{tr}(\Sigma_m^{1/2}) = \sum_{i=1}^p \sigma_i$, where $\sigma_i > 0$ are the values on the diagonal of $\Sigma_m^{1/2}$. Let $\tilde{G} = V_m'G'U_m$ with values $\tilde{g}_{ii}$ on the diagonal, then, as $\Sigma_m^{1/2}$ is a diagonal matrix, it follows that

$$\mathrm{tr}(V_m'G'U_m\Sigma_m^{1/2}) = \mathrm{tr}(\tilde{G}\Sigma_m^{1/2}) = \sum_{i=1}^p \sigma_i \tilde{g}_{ii}.$$

To prove that this is at most $\sum_{i=1}^p \sigma_i$ it suffices to prove that $|\tilde{g}_{ii}| \leq 1$ for all $i = 1, \ldots, p$. Let u_i be the i-th column of U_m and let v_i be the i-th column of GV_m, then $U_m'U_m = I_{p\times p}$ implies that $u_i'u_i = 1$ and $V_m'G'GV_m = I_{p\times p}$ implies that $v_i'v_i = 1$. Therefore, $\tilde{g}_{ii} = v_i'u_i \leq (v_i'v_i)^{1/2}(u_i'u_i)^{1/2} = 1$.

This concludes the proof that $\hat{G} = U_m V_m'$ solves (20).

(f) Summary of algorithm

We summarize the above steps in the following algorithm. As noted in step (a), the algorithm should be applied in an iterative way to X and the residuals $y_{\mathrm{res}} = (y - \alpha - \sum_{j=0}^r Z(-j)\gamma_j)$ that are obtained from the current estimates of $(\alpha, \gamma_0, \ldots, \gamma_r)$.

1. **Initialization**

   Construct initial estimates $(\overline{A}, \overline{B}, \overline{\beta})$, with $\overline{A}'X'X\overline{A} = I_{p\times p}$. For instance, let $X = U_x\Sigma_x V_x'$ be an SVD with Σ_x a square, invertible matrix (we assume that X has full column rank, but the steps below are easily adjusted in case X has reduced column rank). Further let V_p consist of the first p columns of V_x and let Σ_p be the $p \times p$ diagonal matrix with the p largest singular values of X on the diagonal. Define $\overline{A} = V_p\Sigma_p^{-1}$, then $\overline{A}'X'X\overline{A} = I_{p\times p}$ and the factors $F = X\overline{A}$ consist of the first p principal components of X. Further define $\overline{B} = \overline{A}'X'X = \Sigma_p V_p'$ and define $\overline{\beta}$ by regressing y on $S_0X\overline{A}, \ldots, S_qX\overline{A}$, so that $\overline{\beta}_j$ is the vector of coefficients belonging to the regressor sub-vector $S_jX\overline{A}$. Define $H_+ = (X'X)^{1/2} = V_x\Sigma_x V_x'$, $H_- = (X'X)^{-1/2} = V_x\Sigma_x^{-1}V_x'$, and $\lambda = (q+1)$.

2. **Computation**

   Compute $\overline{c}_j = X\overline{A}\overline{\beta}_j$ and $u_j = \lambda\overline{c}_j + S_j'(y_{\mathrm{res}} - R\overline{c})$ for $j = 0, \ldots, q$, and compute $V = (w_1/\lambda)X'(\sum u_j u_j')X + w_2X'XX'X$ and $W = H_-VH_-$.

3. **Update of A**

   Compute $\overline{G} = H_+\overline{A}$ and $M = \overline{G}'W^2\overline{G}$, compute an eigendecomposition of the $p \times p$

matrix M, say $M = V_m \Sigma_m V_m'$, compute $U_m = W\overline{G}V_m\Sigma_m^{-1/2}$, and update G by $\hat{G} = U_m V_m'$. The updated estimate of A is defined as $\hat{A} = H_- \hat{G}$.

4. **Update of** (B, β)
   Compute updated estimates of B and β by $\hat{B} = \hat{A}'X'X = \hat{G}'H_+$ and $\hat{\beta}_j = (1/\lambda)\hat{A}'X'u_j$, $j = 0, \ldots, q$.

5. **Update of** $(\alpha, \gamma_0, \ldots, \gamma_r)$
   Use the estimates $(\hat{A}, \hat{B}, \hat{\beta})$ of Steps 3 and 4 to update the estimates of $(\alpha, \gamma_0, \ldots, \gamma_r)$ in (7) by regressing the residuals $(y - \sum_{j=0}^{q} S_j X \hat{A}\hat{\beta}_j)$ on a constant and Z and its r lags. Compute the corresponding updated residuals $y_{\text{res}} = (y - \hat{\alpha} - \sum_{j=0}^{r} Z(-j)\hat{\gamma}_j)$.

6. **Iteration**
   Return to Step 2, using the residuals y_{res} of Step 5 and replacing $(\overline{A}, \overline{\beta})$ by $(\hat{A}, \hat{\beta})$ of Steps 3 and 4. Iterate Steps 2 to 5 until the PCovR criterion values in (7) converge.

In Step 6 we used a stopping criterion in terms of the relative improvement of the PCovR criterion (7). For instance, in the empirical applications in Section 5, we stopped the iterations if $(f_p - f_c)/f_p < 10^{-6}$, where f_p and f_c denote respectively the previous and current value of the PCovR criterion (7).

Appendix A.2 (Section 3.4)
Non-convexity of the PCovR criterion function

We prove the assertion made at the end of Section 3.4 that the function

$$f(A, B, \alpha, \beta) = w_1||y - \alpha - XA\beta||^2 + w_2||X - XAB||^2$$

in (4) is not convex in its arguments $\theta = (A, B, \alpha, \beta)$. For this purpose it suffices to construct data (y, X), two parameter sets θ_1 and θ_2, and a scalar value $0 < h < 1$ so that

$$f(h\theta_1 + (1-h)\theta_2) > hf(\theta_1) + (1-h)f(\theta_2).$$

Let $\theta_1 = (A, 0, 0, b)$ with $B = 0$ and $\alpha = 0$ and with $Ab \neq 0$, and let $\theta_2 = (0, 0, 0, 0)$. We consider (special) data with $y = XAb \neq 0$, so that θ_1 provides a prefect fit of y. It follows that $f(\theta_1) = w_2||X||^2$ and $f(\theta_2) = w_1||y||^2 + w_2||X||^2$, so that

$$hf(\theta_1) + (1-h)f(\theta_2) = (1-h)w_1||y||^2 + w_2||X||^2.$$

As $h\theta_1 + (1-h)\theta_2 = (hA, 0, 0, hb)$ we get

$$f(h\theta_1 + (1-h)\theta_2) = w_1||y - X(hA)(hb)||^2 + w_2||X||^2 = w_1(1-h^2)^2||y||^2 + w_2||X||^2.$$

As $y \neq 0$, it suffices to find a value $0 < h < 1$ so that $(1-h^2)^2 > (1-h)$. This is possible, as $(1-h^2)^2/(1-h) = (1-h^2)(1+h) = 1+h-h^2-h^3 > 1$ for $h > 0$ sufficiently small, for instance, for $h = 0.1$.

This concludes the proof that f is not convex.

Appendix A.3 (Section 3.5)
Approximation of the correlation between X and F

We consider the approximation of the squared correlation $\rho^2_{XF}(p)$ to derive the upper bound (9) in Section 3.5. The argument is based on the fact that PCR, or equivalently SVD, is a consistent method to estimate the factors in a factor model, see for instance Stock and Watson (2002a). We assume for simplicity that all variables are scaled to unit norm, that is, each column x_i of the predictor matrix X has norm 1, $i = 1, \ldots, k$.

Let $X = USV'$ be an SVD of X with ordered singular values $s_1 \geq s_2 \geq \ldots \geq s_k$, where $s_{T+1} = \ldots = s_k = 0$ if $k > T$. As each column of X has norm 1, it follows that $\sum_{i=1}^k s_i^2 = ||X||^2 = k$. Let $\hat{X}(p)$ be the SVD approximation of X of rank p, which provides a consistent estimator of the first p principal factors. Then $||X - \hat{X}(p)||^2 = \sum_{i=p+1}^k s_k^2$. Let $\hat{x}_i(p)$ be the i-th column of $\hat{X}(p)$, then the R-squared between x_i and its approximation $\hat{x}_i(p)$ is

$$R^2_{x_i \hat{x}_i(p)} = 1 - \frac{||x_i - \hat{x}_i(p)||^2}{||x_i||^2} = 1 - ||x_i - \hat{x}_i(p)||^2.$$

If we average this result over the k variables we get

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^k R^2_{x_i \hat{x}_i(p)} &= 1 - \frac{1}{k}\sum_{i=1}^k ||x_i - \hat{x}_i(p)||^2 = 1 - \frac{1}{k}||X - \hat{X}(p)||^2 \\ &= 1 - \frac{\sum_{i=p+1}^k s_i^2}{\sum_{i=1}^k s_i^2} = \frac{\sum_{i=1}^p s_i^2}{\sum_{i=1}^k s_i^2}. \end{aligned}$$

If we take $R^2_{x_i \hat{x}_i(p)}$ as an approximation of $\rho^2_{x_i F}(p)$, then we get

$$\rho^2_{XF}(p) = \frac{1}{k}\sum_{i=1}^k \rho^2_{x_i F}(p) \approx \frac{1}{k}\sum_{i=1}^k R^2_{x_i \hat{x}_i(p)} = \frac{\sum_{i=1}^p s_i^2}{\sum_{i=1}^k s_i^2}.$$

This result provides the approximation $\rho^2_{XF}(p) - \rho^2_{XF}(p-1) \approx s_p^2 / \sum_{i=1}^k s_i^2$ used in Section 3.5.

Finally we show that, if we employ the same idea with more factors to fit y, this does not lead to a tighter bound for w. Suppose that $p > 1$ and that we use $p_m \leq p$ factors to fit y, where $kp_m \leq T$, and that we use the remaining $(p - p_m)$ factors to fit X. Then by arguments similar to the ones used in Section 3.5 for the case $p_m = 1$, it follows that we

can create (at least) kp_m zero errors in fitting y and that, in order to prevent overfitting, we should require that

$$w\Big(1-\frac{kp_m}{T}\Big)+(1-w)\Big(1-\rho^2_{XF}(p-p_m)\Big) > w(1-\rho^2_{yF})+(1-w)\Big(1-\rho^2_{XF}(p)\Big),$$

or equivalently,

$$w < \frac{\rho^2_{XF}(p)-\rho^2_{XF}(p-p_m)}{(kp_m/T)+(\rho^2_{XF}(p)-\rho^2_{XF}(p-p_m))}.$$

The above arguments can be used again to approximate $\rho^2_{XF}(p)-\rho^2_{XF}(p-p_m)$ by means of $\sum_{i=p-p_m+1}^{p} s_i^2 / \sum_{i=1}^{k} s_i^2$. This gives the bound

$$w < \frac{\sum_{i=p-p_m+1}^{p} s_i^2 / \sum_{i=1}^{k} s_i^2}{(kp_m/T)+\sum_{i=p-p_m+1}^{p} s_i^2 / \sum_{i=1}^{k} s_i^2} = \frac{(\sum_{i=p-p_m+1}^{p} s_i^2/p_m) / \sum_{i=1}^{k} s_i^2}{(k/T)+(\sum_{i=p-p_m+1}^{p} s_i^2/p_m) / \sum_{i=1}^{k} s_i^2}.$$

Because $s_i^2 \geq s_p^2$ for all $i = p-p_m+1,\ldots,p$ it follows that $\sum_{i=p-p_m+1}^{p} s_i^2/p_m \geq s_p^2$, and this implies that the above bound is larger than or equal to (9) in Section 3.5.

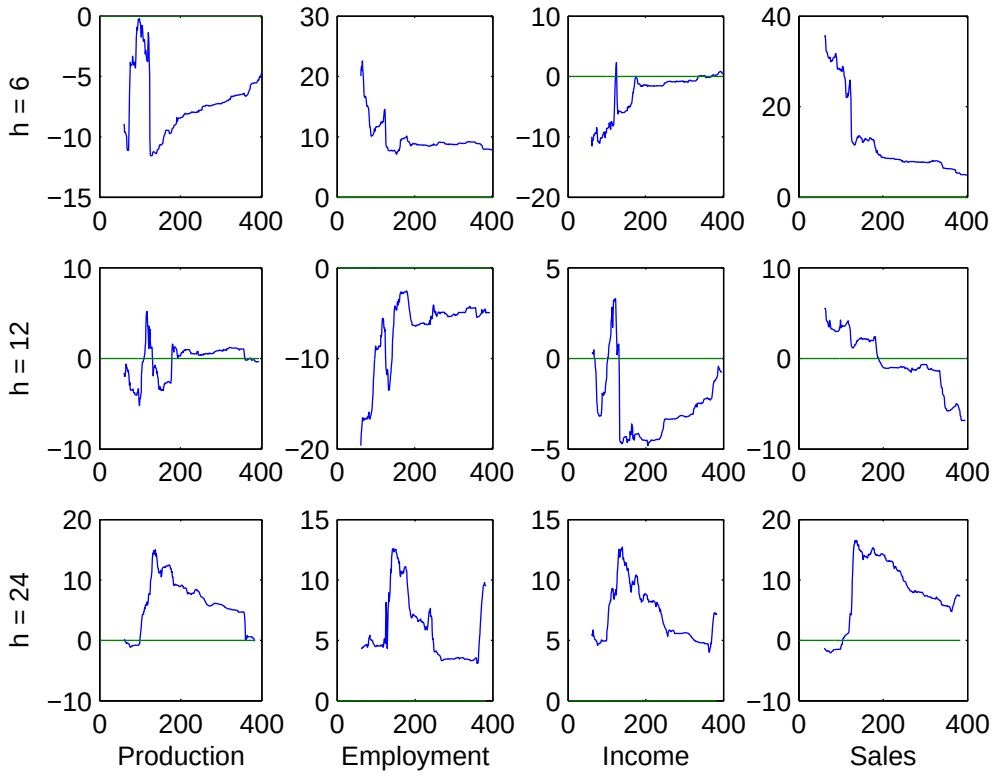

Figure 1: Percentage MSE gain of PCovR as compared to PCR. The gain at time t is the percentage gain in MSE of PCovR as compared to PCR, both using multiple factor models, when evaluated over the forecast interval starting at 1970.01 ($t = 1$) and ending at varying times, ranging from 1975.01 ($t = 61$) till the end of the sample ($t = 408 - h$).

Table 1: MSE of PCovR for various DGP's and forecast models.

					p	1								2							
					L	1				5				1				5			
k	L_1	L_2	ρ_y^2	rvar(y)	w	10^{-4}	0.1	0.5	0.9	10^{-4}	0.1	0.5	0.9	10^{-4}	0.1	0.5	0.9	10^{-4}	0.1	0.5	0.9
10	1	-1	0.5	2		1.99	1.53	*1.19*	*1.19*	2.18	1.79	1.34	1.34	1.88	1.43	1.26	1.36	2.13	1.80	1.66	1.70
10	1	-1	0.9	10		9.34	2.44	1.25	*1.23*	10.08	3.32	1.37	1.35	8.21	2.09	1.27	1.31	9.17	2.93	1.49	1.68
10	5	-1	0.5	2		2.29	2.30	2.50	2.57	2.31	1.96	1.46	*1.45*	2.31	2.35	2.66	2.81	2.28	1.95	1.79	1.87
10	5	-1	0.9	10		10.04	10.18	11.10	11.30	9.20	2.94	*1.31*	*1.31*	10.16	10.39	11.80	12.29	8.54	2.71	1.46	1.67
10	1	1	0.5	3		2.09	1.74	*1.24*	*1.24*	2.30	2.00	1.42	1.40	1.94	1.58	1.29	1.38	2.31	1.94	1.67	1.76
10	1	1	0.9	19		8.98	4.99	1.24	*1.19*	10.08	6.14	1.47	1.39	7.94	3.67	1.26	1.26	9.90	5.12	1.63	1.81
10	1	5	0.5	3		2.86	2.56	2.19	2.21	2.28	2.01	*1.37*	*1.37*	2.75	2.36	2.22	2.33	2.32	1.94	1.68	1.77
10	1	5	0.9	19		17.20	13.01	10.73	10.80	10.40	6.17	1.33	*1.24*	16.00	11.78	11.32	11.73	9.81	5.07	1.43	1.60
10	5	1	0.5	3		2.06	2.08	2.28	2.37	2.16	1.81	*1.25*	1.27	2.10	2.13	2.37	2.55	2.15	1.80	1.55	1.65
10	5	1	0.9	19		10.76	10.78	11.31	11.89	10.71	6.38	1.39	*1.30*	10.85	10.93	11.74	12.61	10.35	5.52	1.51	1.67
10	5	5	0.5	3		2.89	2.92	3.31	3.32	2.17	1.96	1.37	*1.35*	2.94	3.03	3.46	3.61	2.21	1.98	1.64	1.74
10	5	5	0.9	19		17.93	18.09	20.22	20.61	10.37	6.20	1.42	*1.34*	18.23	18.68	21.25	21.73	9.80	5.12	1.50	1.67
40	1	-1	0.5	2		1.99	*1.73*	1.89		2.19	2.55	3.00		1.94	1.94	5.23		2.20	3.46	4.13	
40	1	-1	0.9	10		10.07	2.10	*1.67*		10.95	3.71	3.87		9.53	2.14	3.33		10.75	5.10	6.47	
40	5	-1	0.5	2		2.29	3.12	3.77		2.39	2.77	3.08		*2.32*	4.88	9.82		2.40	3.53	4.26	
40	5	-1	0.9	10		10.45	16.30	19.12		10.78	*4.05*	4.07		10.61	25.02	48.68		10.60	4.80	6.07	
40	1	1	0.5	3		1.99	*1.72*	1.86		2.23	2.60	3.20		1.96	1.87	4.58		2.34	3.28	4.25	
40	1	1	0.9	19		9.77	3.09	*1.77*		11.24	5.66	4.16		9.29	3.07	2.13		11.24	6.49	6.65	
40	1	5	0.5	3		2.82	3.16	3.69		*2.21*	2.26	2.74		2.85	4.07	9.68		2.28	3.06	3.96	
40	1	5	0.9	19		17.73	16.26	18.03		11.08	5.88	*4.46*		17.58	18.62	48.09		11.39	6.53	6.24	
40	5	1	0.5	3		*2.01*	2.56	3.68		2.10	2.38	2.81		2.07	3.20	9.49		2.27	3.20	4.10	
40	5	1	0.9	19		10.13	12.37	19.18		10.64	5.48	*4.29*		10.38	13.96	45.35		11.25	6.95	6.84	
40	5	5	0.5	3		2.89	4.32	5.50		*2.22*	2.69	3.21		2.90	6.92	14.52		2.26	3.28	4.43	
40	5	5	0.9	19		17.83	27.71	33.64		10.96	5.50	*4.08*		18.27	45.89	94.60		11.03	6.58	6.40	
100	1	-1	0.5	2		2.36	**			9.44	**			*2.14*	**			7.01	**		
100	1	-1	0.9	10		10.21	**			14.16	**			*10.18*	**			14.06	**		
100	5	-1	0.5	2		2.24	**			*2.16*	**			5.22	**			4.76	**		
100	5	-1	0.9	10		*14.75*	**			21.38	**			22.34	**			29.36	**		
100	1	1	0.5	3		*2.07*	**			8.78	**			6.85	**			42.98	**		
100	1	1	0.9	19		10.19	**			35.44	**			*9.98*	**			44.28	**		
100	1	5	0.5	3		7.74	**			*2.69*	**			16.74	**			4.92	**		
100	1	5	0.9	19		18.74	**			27.39	**			28.00	**			*13.45*	**		
100	5	1	0.5	3		*2.15*	**			2.60	**			12.05	**			9.42	**		
100	5	1	0.9	19		*10.98*	**			14.61	**			13.25	**			12.07	**		
100	5	5	0.5	3		3.77	**			*2.26*	**			3.71	**			2.40	**		
100	5	5	0.9	19		18.70	**			*11.43*	**			*	**			58.17	**		

The first five DGP columns show the number of predictors k, the lags L_1 of the DGP predictor x^* and L_2 of the preferential predictor z, the squared correlation $\rho_y^2 = \rho_{yx}^2 = \rho_{yz}^2$ between y and x and between y and z, and the relative variance of y, rvar(y). The sixteen MSE columns show the MSE defined in (14) for PCovR in forecast models with $p = 1$ or 2 factors, with equal lags $L = q = r = 1$ or 5, and with PCovR weight $w = 0.0001$, 0.1, 0.5 or 0.9. For each DGP row, values in italics show the methods with the smallest MSE. A * stands for MSE values between 10^2 and 10^4 and ** for values larger than 10^4.

Table 2: MSE of PCovR when the forecast model is selected by BIC or CV.

						BIC					CV				
k	L_1	L_2	ρ_y^2	rvar(y)	w	10^{-4}	0.1	0.5	0.9	BIC	10^{-4}	0.1	0.5	0.9	CV
10	1	-1	0.5	2		1.99	1.39	1.29	1.35	1.99	1.74	1.30	*1.14*	*1.14*	1.15
10	1	-1	0.9	10		7.97	1.82	1.23	1.31	7.97	6.29	1.69	1.22	*1.20*	*1.20*
10	5	-1	0.5	2		2.23	2.24	2.36	2.41	2.23	1.90	1.58	*1.27*	1.29	1.29
10	5	-1	0.9	10		9.95	8.09	2.35	1.89	9.95	8.10	2.22	*1.25*	1.26	1.26
10	1	1	0.5	3		2.12	1.60	1.32	1.40	2.12	1.76	1.47	*1.26*	1.27	1.27
10	1	1	0.9	19		7.78	3.09	1.27	1.32	7.78	6.78	3.02	1.29	*1.24*	*1.24*
10	1	5	0.5	3		2.18	1.65	*1.29*	1.39	2.18	1.95	1.64	1.35	1.36	1.37
10	1	5	0.9	19		8.41	3.16	*1.19*	1.21	8.41	7.29	3.46	1.41	1.35	1.38
10	5	1	0.5	3		2.06	2.08	2.16	2.22	2.06	1.91	1.63	1.30	*1.27*	1.28
10	5	1	0.9	19		10.82	10.27	3.23	2.08	10.82	8.70	4.25	1.34	*1.24*	1.26
10	5	5	0.5	3		2.24	2.25	2.39	2.46	2.24	2.09	1.79	1.38	*1.40*	1.42
10	5	5	0.9	19		11.41	11.09	3.32	1.95	11.41	9.38	4.52	1.35	*1.31*	*1.31*
40	1	-1	0.5	2		2.01	2.81	5.71		2.01	1.94	*1.70*	1.93		1.86
40	1	-1	0.9	10		10.23	3.36	4.35		10.23	9.47	2.22	*1.87*		1.91
40	5	-1	0.5	2		2.23	2.94	5.77		2.23	2.06	2.43	3.21		*1.98*
40	5	-1	0.9	10		10.32	14.35	28.53		10.32	10.31	*3.03*	3.42		3.42
40	1	1	0.5	3		2.07	2.68	5.42		2.07	*1.95*	*1.95*	2.10		*1.95*
40	1	1	0.9	19		10.19	6.92	2.70		10.19	9.87	3.57	*1.98*		*1.98*
40	1	5	0.5	3		2.22	2.99	4.91		2.22	2.19	*2.05*	2.79		*2.05*
40	1	5	0.9	19		10.93	9.25	2.59		10.93	9.84	3.89	*2.00*		2.02
40	5	1	0.5	3		*2.03*	2.47	6.10		*2.03*	2.21	2.38	3.50		2.16
40	5	1	0.9	19		10.29	11.85	27.74		10.29	11.00	5.46	*3.83*		4.01
40	5	5	0.5	3		*2.17*	2.91	6.88		*2.17*	2.35	2.77	3.61		2.36
40	5	5	0.9	19		10.81	12.98	30.73		10.81	10.99	6.13	*4.01*		4.77
100	1	-1	0.5	2		*1.97*					2.19				
100	1	-1	0.9	10		10.45					*10.05*				
100	5	-1	0.5	2		5.44					*2.08*				
100	5	-1	0.9	10		10.77					*9.99*				
100	1	1	0.5	3		*2.08*					2.24				
100	1	1	0.9	19		10.32					*10.06*				
100	1	5	0.5	3		*2.15*					2.22				
100	1	5	0.9	19		10.96					*10.86*				
100	5	1	0.5	3		2.07					*2.01*				
100	5	1	0.9	19		10.36					*10.26*				
100	5	5	0.5	3		*2.22*					2.28				
100	5	5	0.9	19		11.06					*10.51*				

The first five DGP columns specify the DGP, see Table 1. The PCovR forecast model is selected by BIC or CV, for fixed weight w and also for w selected by BIC or CV. For each DGP row, values in italics show the methods with the smallest MSE.

Table 3: Structure of PCovR forecast model, selected by BIC or CV.

					p					q					r					w
					BIC		CV			BIC		CV			BIC		CV			CV
k	L_1	L_2	ρ_y^2	w	10^{-4}	0.1	10^{-4}	0.1	CV	10^{-4}	0.1	10^{-4}	0.1	CV	10^{-4}	0.1	10^{-4}	0.1	CV	CV
10	1	-1	0.5		3.00	2.98	2.30	2.21	1.23	0.12	0.88	1.20	1.23	1.57	-0.90	-0.92	0.24	-0.27	-0.17	0.66
10	1	-1	0.9		3.00	2.93	2.52	2.50	1.24	0.47	1.00	1.19	1.07	1.70	-0.61	-0.92	0.43	-0.47	-0.23	0.78
10	5	-1	0.5		3.00	3.00	1.76	1.49	1.11	0.00	0.00	2.71	4.31	4.90	-0.95	-0.94	-0.30	-0.41	-0.22	0.64
10	5	-1	0.9		3.00	2.99	2.06	1.69	1.05	0.00	1.03	3.69	5.00	5.00	-0.97	-0.96	-0.16	-0.86	-0.41	0.80
10	1	1	0.5		3.00	2.99	2.27	2.29	1.24	0.14	0.70	1.01	1.17	1.52	1.20	1.23	1.81	1.68	1.66	0.69
10	1	1	0.9		2.99	2.96	2.53	2.65	1.20	0.48	0.98	1.01	1.01	1.60	1.48	1.42	2.02	1.66	1.60	0.85
10	1	5	0.5		2.98	2.99	2.20	2.23	1.24	0.13	0.66	1.03	1.12	1.34	4.89	4.98	4.98	4.98	4.99	0.67
10	1	5	0.9		2.99	2.97	2.46	2.62	1.21	0.46	0.98	1.06	1.07	1.50	5.00	5.00	5.00	5.00	5.00	0.85
10	5	1	0.5		3.00	3.00	1.60	1.58	1.11	0.00	0.00	2.16	3.90	4.91	1.02	1.03	1.53	1.41	1.58	0.67
10	5	1	0.9		3.00	3.00	1.84	2.09	1.06	0.00	0.18	2.94	4.84	5.00	1.02	1.02	1.35	1.12	1.46	0.86
10	5	5	0.5		2.99	3.00	1.63	1.60	1.15	0.00	0.00	2.54	3.89	4.79	4.81	4.80	4.96	4.98	4.98	0.64
10	5	5	0.9		2.99	3.00	1.86	2.08	1.06	0.00	0.17	3.34	4.86	5.00	4.99	4.99	5.00	5.00	5.00	0.86
40	1	-1	0.5		3.00	2.89	1.76	1.10	1.44	0.00	0.00	1.13	1.19	1.16	-0.92	0.05	0.01	-0.02	-0.06	0.08
40	1	-1	0.9		3.00	2.85	2.03	1.49	1.07	0.00	0.91	1.13	1.09	1.25	-0.95	-0.63	0.12	-0.69	-0.37	0.46
40	5	-1	0.5		3.00	2.89	1.43	1.33	1.39	0.00	0.00	1.16	3.35	1.55	-0.96	-0.29	-0.32	-0.05	-0.30	0.02
40	5	-1	0.9		2.99	2.91	1.60	1.25	1.25	0.00	0.00	1.53	4.94	4.64	-0.97	-0.51	-0.39	-0.60	-0.42	0.35
40	1	1	0.5		2.99	2.91	1.79	1.25	1.45	0.00	0.00	0.87	1.03	1.03	1.16	1.83	1.87	1.68	1.72	0.09
40	1	1	0.9		2.99	2.88	2.01	1.83	1.08	0.00	0.52	0.88	1.02	1.14	1.17	1.52	2.02	1.28	1.44	0.49
40	1	5	0.5		2.98	2.90	1.78	1.28	1.55	0.00	0.00	0.92	1.17	1.06	4.84	4.14	4.93	4.81	4.94	0.05
40	1	5	0.9		2.98	2.88	1.96	1.73	1.15	0.00	0.42	0.94	1.04	1.18	5.00	4.92	5.00	5.00	5.00	0.48
40	5	1	0.5		2.99	2.91	1.43	1.25	1.39	0.00	0.00	0.86	2.44	1.49	1.05	1.52	1.43	1.73	1.49	0.03
40	5	1	0.9		3.00	2.94	1.46	1.23	1.21	0.00	0.00	1.04	4.33	4.38	1.02	1.30	1.33	1.24	1.37	0.34
40	5	5	0.5		2.98	2.91	1.41	1.26	1.38	0.00	0.00	1.22	2.97	1.79	4.82	3.94	4.91	4.85	4.92	0.03
40	5	5	0.9		2.97	2.94	1.46	1.25	1.23	0.00	0.00	1.39	4.28	4.21	5.00	4.75	5.00	5.00	5.00	0.33
100	1	-1	0.5		3.00		1.65			0.00		1.03			-0.92		-0.28			
100	1	-1	0.9		3.00		1.81			0.00		1.02			-0.92		-0.19			
100	5	-1	0.5		3.00		1.47			0.00		0.86			-0.95		-0.47			
100	5	-1	0.9		3.00		1.45			0.00		1.00			-0.98		-0.56			
100	1	1	0.5		2.98		1.57			0.00		0.67			1.18		1.77			
100	1	1	0.9		2.98		1.77			0.00		0.73			1.15		1.80			
100	1	5	0.5		2.98		1.61			0.00		0.68			4.89		4.95			
100	1	5	0.9		2.98		1.75			0.00		0.75			5.00		5.00			
100	5	1	0.5		2.99		1.45			0.00		0.69			1.05		1.43			
100	5	1	0.9		2.99		1.40			0.00		0.60			1.03		1.26			
100	5	5	0.5		2.97		1.39			0.00		0.89			4.86		4.95			
100	5	5	0.9		2.99		1.42			0.00		0.88			5.00		4.99			

The first five DGP columns specify the DGP, see Table 1. The other columns show the mean values for the selected BIC and CV models of the number of factors p, factor lags q (true value is L_1), and preferential predictor lags r (true value is L_2), and also the mean PCovR weight w selected by CV (from the set $\{10^{-4}, 0.01, 0.1, 0.5, 0.9\}$).

Table 4: MSE of PCR and PCovR with single factor, for four variables and three horizons.

	method			gain			
	PCR	PCovR	PCovR∗	70-03	70-80	81-91	92-03
$h = 6$							
Production	94.9	74.4	73.0	21.6	32.8	-10.4	18.8
Employment	93.0	78.0	75.0	16.2	13.6	29.9	-4.5
Income	84.1	77.1	77.8	8.3	26.6	-33.7	5.0
Sales	96.6	69.2	64.0	28.4	35.3	28.2	-5.2
Average	92.2	74.7	72.4	18.6	27.1	3.5	3.5
$h = 12$							
Production	98.4	66.0	55.9	32.9	45.5	20.0	7.5
Employment	91.5	74.9	67.5	18.1	14.5	30.3	2.8
Income	89.0	80.8	73.7	9.2	22.5	-14.9	4.4
Sales	97.4	64.8	57.4	33.4	40.1	36.5	-11.8
Average	94.1	71.6	63.6	23.4	30.7	18.0	0.7
$h = 24$							
Production	101.7	62.2	54.1	38.8	54.9	33.4	9.1
Employment	98.2	78.2	70.3	20.3	31.5	8.3	13.1
Income	97.6	85.6	78.8	12.3	18.8	-2.0	14.1
Sales	98.6	65.3	52.0	33.7	37.5	22.8	30.3
Average	99.0	72.8	63.8	26.3	35.7	15.6	16.6

The columns PCR and PCovR show the MSE of respectively PCR and PCovR (with CV weight), as percentage of the MSE of the AR benchmark model, and the column PCovR∗ shows this percentage MSE for a posteriori optimal choice of a fixed weight w. The four gain columns show the percentage gain (+) or loss (-) of the MSE of PCovR (with CV weight) as compared to PCR, over the full sample and over three sub-samples.

Table 5: MSE of PCR and PCovR with multiple factors, for four variables and three horizons.

	method			gain			
	PCR	PCovR	PCovR∗	70-03	70-80	81-91	92-03
$h = 6$							
Production	64.2	67.3	65.8	-4.7	-11.3	-1.1	6.6
Employment	86.2	79.4	75.4	7.9	7.8	12.4	-3.3
Income	73.4	73.0	71.9	0.5	-5.7	5.4	4.7
Sales	67.1	63.8	62.3	4.9	11.6	-0.5	-10.4
Average	72.7	70.9	68.8	2.1	0.6	4.0	-0.6
$h = 12$							
Production	54.5	54.7	55.2	-0.4	-0.6	1.8	-2.7
Employment	67.0	70.4	65.1	-5.0	-12.6	5.3	-5.3
Income	67.8	68.3	67.7	-0.7	-0.9	-6.1	6.3
Sales	46.4	49.6	46.5	-6.8	1.3	-5.7	-27.7
Average	58.9	60.7	58.6	-3.2	-3.2	-1.2	-7.3
$h = 24$							
Production	55.5	55.6	51.3	-0.1	14.3	-13.8	-11.2
Employment	71.8	65.0	68.5	9.5	8.2	-1.2	27.2
Income	75.3	69.9	72.4	7.1	12.3	-3.3	10.6
Sales	54.6	50.6	50.6	7.4	15.9	-17.9	1.8
Average	64.3	60.3	60.7	6.0	12.7	-9.0	7.1

This table is similar to Table 4, but now for multiple factor models with number of factors ($p \leq 4$) selected by BIC.

Table 6: Statistics of models and forecast errors for PCR and PCovR with multiple factors.

	models							forecast errors					
	PCR			PCovR				PCR			PCovR		
	p	q	r	p	q	r	w	mean	mabs	std	mean	mabs	std
$h = 6$													
Production	2.16	0.39	0.12	1.74	0.54	0.25	0.29	-0.52	3.33	4.31	-0.33	3.40	4.43
Employment	2.46	0.63	-0.32	1.76	0.72	-0.38	0.40	-0.17	1.26	1.74	-0.10	1.21	1.68
Income	2.32	0.47	-0.76	1.90	0.74	-0.82	0.39	-0.37	2.06	2.58	-0.21	2.03	2.59
Sales	2.29	0.16	0.16	1.80	0.50	1.09	0.29	-0.55	3.26	4.45	-0.49	3.28	4.35
Average	2.31	0.41	-0.20	1.80	0.63	0.03	0.34	-0.40	2.48	3.27	-0.28	2.48	3.26
$h = 12$													
Production	2.31	1.13	0.19	2.21	1.02	-0.72	0.25	-0.86	2.66	3.30	-0.79	2.63	3.32
Employment	2.48	0.98	-0.21	2.00	0.88	-0.44	0.36	-0.35	1.19	1.55	-0.25	1.22	1.61
Income	2.42	0.66	-0.60	2.21	0.74	-0.72	0.25	-0.53	1.74	2.07	-0.48	1.73	2.09
Sales	2.31	1.14	1.04	2.14	1.24	1.32	0.17	-0.70	2.20	2.83	-0.71	2.34	2.92
Average	2.38	0.98	0.11	2.14	0.97	-0.14	0.26	-0.61	1.95	2.44	-0.56	1.98	2.49
$h = 24$													
Production	2.55	1.04	-0.88	2.09	1.16	-0.87	0.32	-0.89	2.20	2.61	-0.84	2.18	2.63
Employment	2.24	1.17	-0.54	2.18	1.11	-0.65	0.31	-0.46	1.18	1.44	-0.38	1.12	1.39
Income	2.64	0.75	-1.00	2.45	0.57	-0.99	0.33	-0.56	1.47	1.74	-0.47	1.41	1.70
Sales	2.66	0.99	0.07	2.01	1.21	0.01	0.45	-0.79	1.84	2.31	-0.73	1.84	2.24
Average	2.52	0.99	-0.59	2.18	1.02	-0.62	0.35	-0.67	1.67	2.03	-0.61	1.64	1.99

Results for multiple factor models selected by BIC, for PCR and PCovR. Shown are the average number of factors p, factor lags q, AR lags r, and (for PCovR) weights w of the selected models. Further, 'mean' is the average forecast error, 'mabs' the average absolute forecast error, and 'std' the standard deviation of the forecast errors.